Ocean Protocol Data Challenge

Forecasting Carbon Emissions Across Continents

Done By : Colin Chan

Table of Contents

Exploratory Data Analysis **4**

Data Cleaning 4

CO2 vs GHG Correlation 5

Sector Contribution Correlation **7**

Emissions vs GDP Correlation **10**

Emissions by Substances Correlation **13**

Temporal Analysis **13**

Prediction Model **18**

Methodology: Linear Regression 18

Data Cleaning / Preparation 18

Linear Regression 19

Results (Test Data Prediction) 19

Methodology: XG Boost 24

Data Cleaning / Preparation 24

Accuracy 26

Introduction

Climate change is one of the most critical challenges facing our planet today. The consequences of rising carbon emissions and their impact on the environment are undeniable, affecting ecosystems, weather patterns, and overall global sustainability. In this report, I analyze data made open by EDGAR - Emissions Database for Global Atmospheric Research European Commission to determine the causes and prevalence of greenhouse gas and carbon dioxide emissions. The objective of this paper is to present insightful findings that shed light into the current state and provide a foundation for more informed decision making using the data provided.

To begin with, an in depth exploratory data analysis was performed to provide further insights on each countries and continent's statistical measures, patterns and trends. This concludes with a prediction model to forecast future emissions using a linear regression and extreme gradient boosting model.

Exploratory Data Analysis

Data Cleaning

1. **Nan values**: NaN values were present and rows with all NaNs were removed from the dataframe. Subsequently, any rows with NaN values within will be replaced with 0.
2. **Cleaning the commas**: Given that the figures were in strings, we adopted the replace(‘,’, ‘’) function to remove the commas and converted these figures into float type.
3. **Missing GDP values before 1990**: Since GDP figures were only available from 1990, any analysis which requires GDP figures where the years are not available will use a truncated dataframe from 1990. However, for the prediction model, given that the amount of data is essential, we included the whole dataframe and treated GDP as 0.

Outlier Analysis for Exploratory Data Analysis

For the first section on exploratory data analysis, we noticed outliers that were present and charted these residuals beyond the boxplot distribution to visualize these.

As for the prediction model, we used the MinMax scaler to scale the data and reduce the impact of outliers. The rationale for these are outlined for each methodology - linear regression, XG Boost algorithms in the later sections.

CO2 vs GHG Correlation

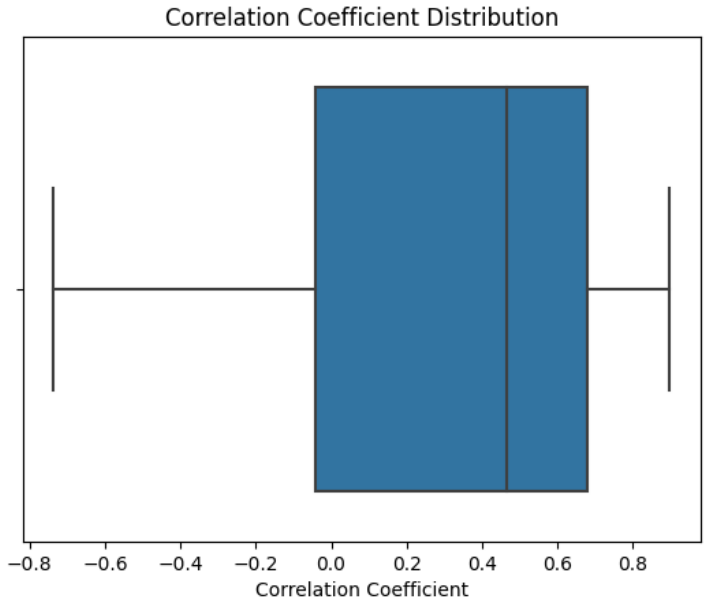
After cleaning the data and merging the dataframes for countries-specific total CO2 emissions and total GHG emissions, the correlations were determined using the function from the numpy package:

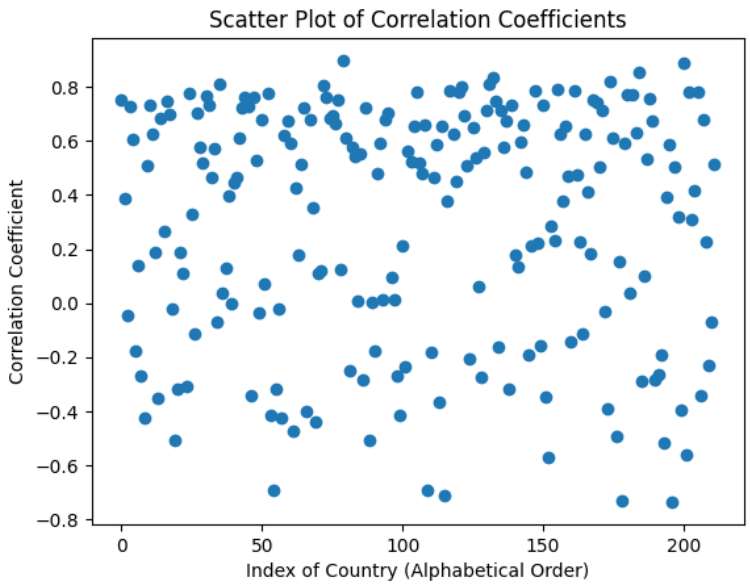
- correlation_matrix = np.corrcoef(fossil_country, ghg_country)
- correlation_coefficient = correlation_matrix[0, 1]

This returned the following results and visualizations to understand the distribution of the correlations.

	Correlation
ABW	0.752092
AFG	0.385693
AGO	-0.047709
AIA	0.728733
AIR	0.605253
...	...
ZAF	0.679704
ZMB	0.226294
ZWE	-0.228792
EU27	-0.068265
GLOBAL TOTAL	0.515958

212 rows × 1 columns





To identify outliers, we used 2 standard deviations from the mean as a base. Should any data point be above this threshold, we will flag them out as outliers. This was done by passing in the stats.zscore method on the correlations .

	index	Correlation
54	DOM	-0.690834
109	LBN	-0.690963
115	LTU	-0.711975
178	SWE	-0.729160
196	UKR	-0.735926

- Dominican Republic
- Lebanon
- Lithuania
- Sweden
- Ukraine

These anomalies all highlight an extremely negative correlation which suggests that the higher the total CO2 emissions, the lower the total GHG emissions and vice versa.

- Sweden is known for its high share of renewable energy, particularly hydropower and wind energy. This could lead to lower CO2 emissions from energy production compared to countries heavily reliant on fossil fuels.
- Ukraine has a diverse industrial sector, including heavy industry. If certain industries in Ukraine have a high carbon intensity, it could contribute to higher CO2 emissions even if the overall GHG emissions are not as high.

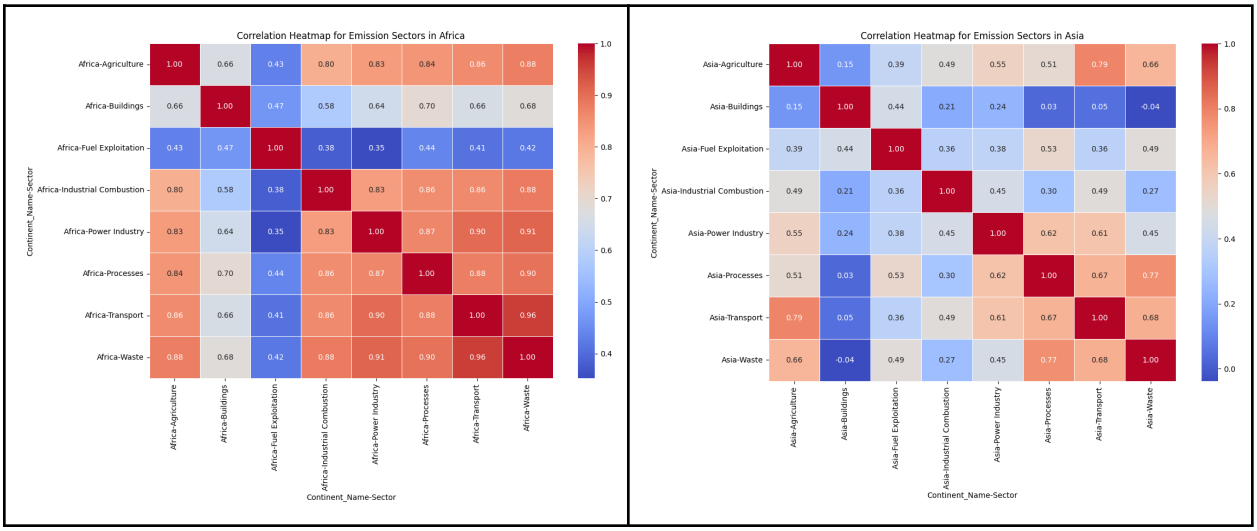
- Lithuania has made efforts to increase its use of renewable energy sources. If Lithuania has successfully transitioned to cleaner energy, it might experience a negative correlation as CO2 emissions decrease while overall GHG emissions remain relatively stable.
- The Dominican Republic might have significant agricultural activities that contribute to GHG emissions, such as methane from livestock or emissions from changes in land use. This could lead to a situation where overall GHG emissions are high, even if CO2 emissions are comparatively lower.

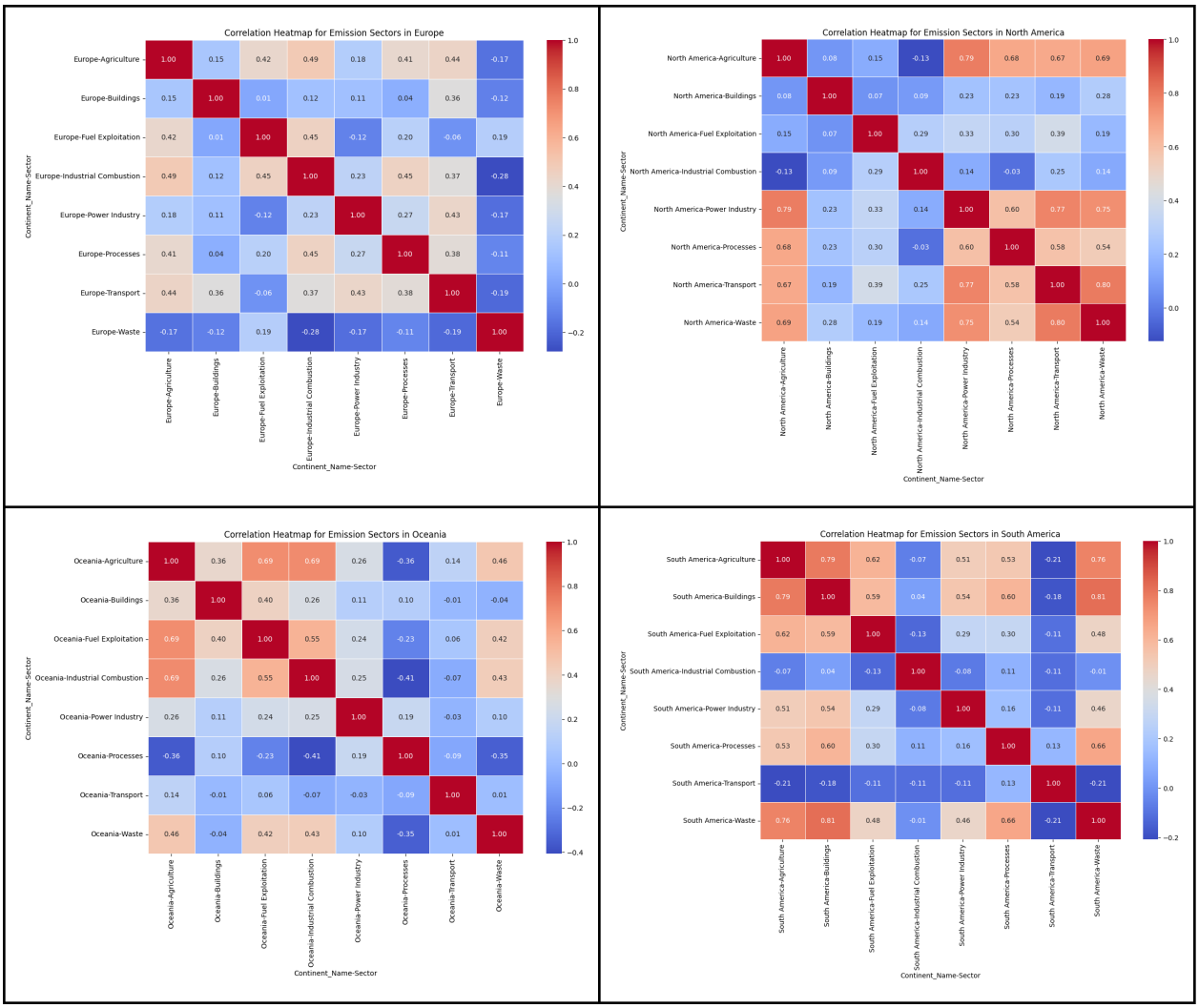
Sector Contribution Correlation

Given that this is organized based on the continent, I used the public dataset (`https://datahub.io/JohnSnowLabs/country-and-continent-codes-list/r/country-and-continent-codes-list-csv.csv`) which maps EDGAR country codes to the respective continent. I then merged the dataframes to retrieve the continent for each country and determined the total sum of emissions aggregated by continent, segmented by sector, on a yearly basis for the metrics - Africa, Asia, Europe, North America, Oceania and South America.

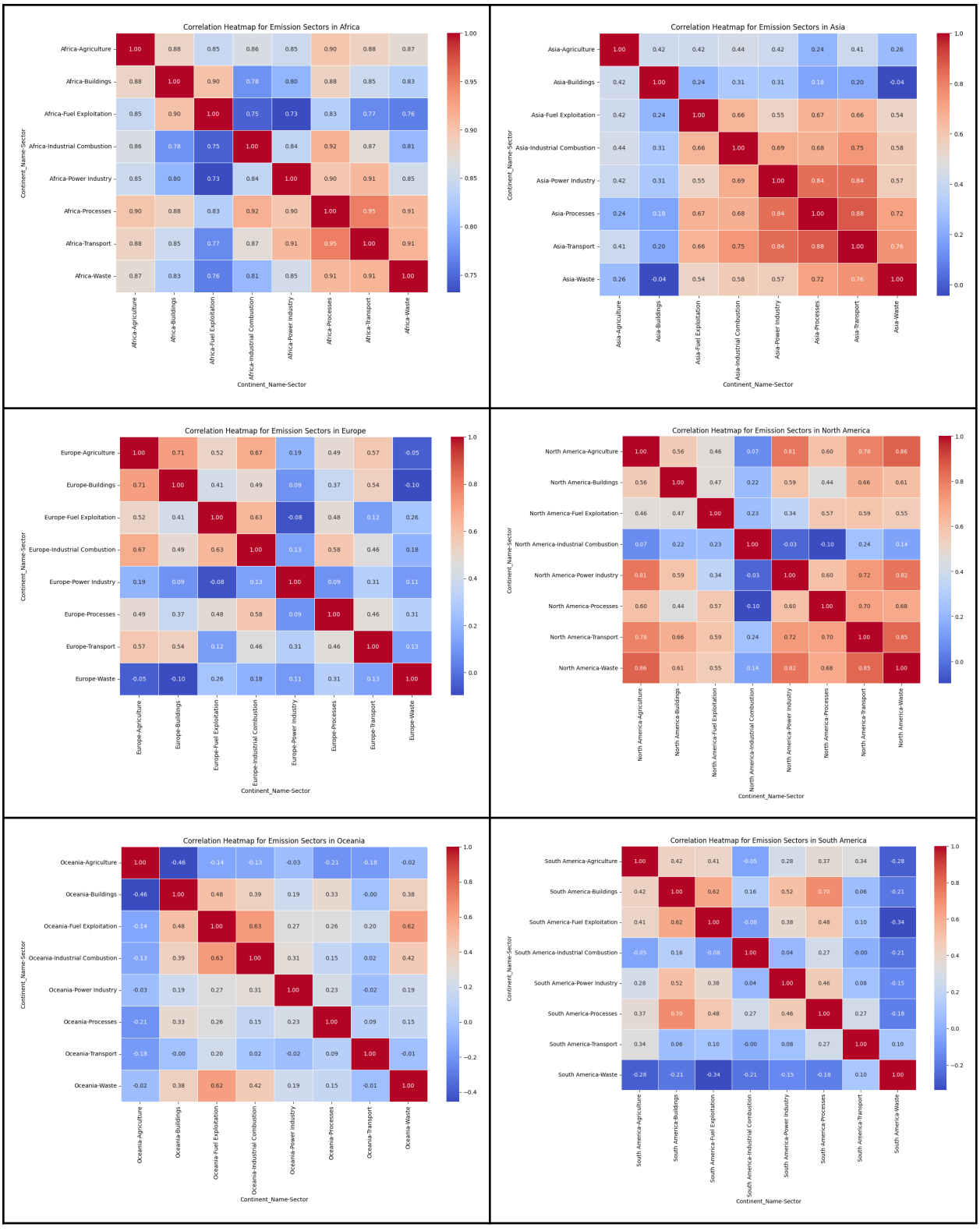
To find the correlation between different sectors, a heatmap was created to visualize the relationship in aggregate emissions between them. This was done based on the continent and type of emissions - CO2, GHG.

CO2 Emissions





GHG Emissions

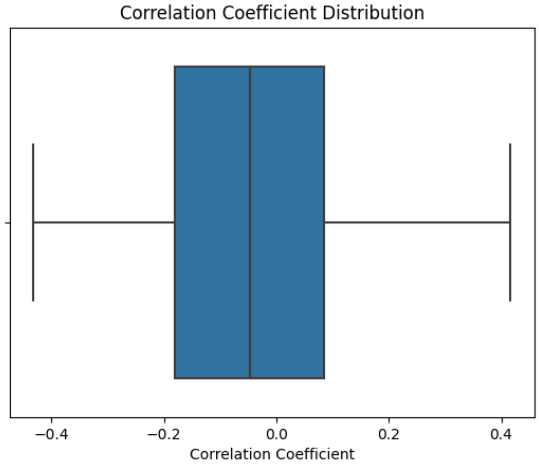

Correlation Heatmap for Emission Sectors in Africa
Africa-Agriculture 1.00 0.88 0.85 0.86 0.85 0.90 0.88 0.87
Africa-Buildings 0.88 1.00 0.90 0.78 0.80 0.88 0.85 0.83
Africa-Fuel Exploitation 0.85 0.90 1.00 0.75 0.73 0.83 0.77 0.76
Africa-Industrial Combustion 0.86 0.78 0.75 1.00 0.84 0.92 0.87 0.81
Africa-Power Industry 0.85 0.80 0.73 0.84 1.00 0.90 0.91 0.85
Africa-Processes 0.90 0.88 0.83 0.92 0.90 1.00 0.95 0.91
Africa-Transport 0.88 0.85 0.77 0.87 0.91 0.95 1.00 0.91
Africa-Waste 0.87 0.83 0.76 0.81 0.85 0.91 0.91 1.00
Continent_Name-Sector
Correlation Heatmap for Emission Sectors in Asia
Asia-Agriculture 1.00 0.42 0.42 0.44 0.42 0.24 0.41 0.26
Asia-Buildings 0.42 1.00 0.24 0.31 0.31 0.18 0.20 -0.04
Asia-Fuel Exploitation 0.42 0.24 1.00 0.66 0.55 0.67 0.66 0.54
Asia-Industrial Combustion 0.44 0.31 0.66 1.00 0.69 0.68 0.75 0.58
Asia-Power Industry 0.42 0.31 0.55 0.69 1.00 0.84 0.84 0.57
Asia-Processes 0.24 0.18 0.67 0.68 0.84 1.00 0.88 0.72
Asia-Transport 0.41 0.20 0.66 0.75 0.84 0.88 1.00 0.76
Asia-Waste 0.26 -0.04 0.54 0.58 0.57 0.72 0.76 1.00
Continent_Name-Sector
Correlation Heatmap for Emission Sectors in Europe
Europe-Agriculture 1.00 0.71 0.52 0.67 0.19 0.49 0.57 -0.05
Europe-Buildings 0.71 1.00 0.41 0.49 0.09 0.37 0.54 -0.10
Europe-Fuel Exploitation 0.52 0.41 1.00 0.63 -0.08 0.48 0.12 0.26
Europe-Industrial Combustion 0.67 0.49 0.63 1.00 0.13 0.58 0.46 0.18
Europe-Power Industry 0.19 0.09 -0.08 0.13 1.00 0.09 0.31 0.11
Europe-Processes 0.49 0.37 0.48 0.58 0.09 1.00 0.46 0.31
Europe-Transport 0.57 0.54 0.12 0.46 0.31 0.46 1.00 0.13
Europe-Waste -0.05 -0.10 0.26 0.18 0.11 0.31 0.13 1.00
Continent_Name-Sector
Correlation Heatmap for Emission Sectors in North America
North America-Agriculture 1.00 0.56 0.46 0.07 0.81 0.60 0.78 0.86
North America-Buildings 0.56 1.00 0.47 0.22 0.59 0.44 0.66 0.61
North America-Fuel Exploitation 0.46 0.47 1.00 0.23 0.34 0.57 0.59 0.55
North America-Industrial Combustion 0.07 0.22 0.23 1.00 -0.03 -0.10 0.24 0.14
North America-Power Industry 0.81 0.59 0.34 -0.03 1.00 0.60 0.72 0.82
North America-Processes 0.60 0.44 0.57 -0.10 0.60 1.00 0.70 0.68
North America-Transport 0.78 0.66 0.59 0.24 0.72 0.70 1.00 0.85
North America-Waste 0.86 0.61 0.55 0.14 0.82 0.68 0.85 1.00
Continent_Name-Sector
Correlation Heatmap for Emission Sectors in Oceania
Oceania-Agriculture 1.00 -0.46 -0.14 -0.13 -0.03 -0.21 -0.18 -0.02
Oceania-Buildings -0.46 1.00 0.48 0.39 0.19 0.33 -0.00 0.38
Oceania-Fuel Exploitation -0.14 0.48 1.00 0.63 0.27 0.26 0.20 0.62
Oceania-Industrial Combustion -0.13 0.39 0.63 1.00 0.31 0.15 0.02 0.42
Oceania-Power Industry -0.03 0.19 0.27 0.31 1.00 0.23 -0.02 0.19
Oceania-Processes -0.21 0.33 0.26 0.15 0.23 1.00 0.09 0.15
Oceania-Transport -0.18 -0.00 0.20 0.02 -0.02 0.09 1.00 -0.01
Oceania-Waste -0.02 0.38 0.62 0.42 0.19 0.15 -0.01 1.00
Continent_Name-Sector
Correlation Heatmap for Emission Sectors in South America
South America-Agriculture 1.00 0.42 0.41 -0.05 0.28 0.37 0.34 -0.28
South America-Buildings 0.42 1.00 0.62 0.16 0.52 0.70 0.06 -0.21
South America-Fuel Exploitation 0.41 0.62 1.00 -0.08 0.38 0.48 0.10 -0.34
South America-Industrial Combustion -0.05 0.16 -0.08 1.00 0.04 0.27 -0.00 -0.21
South America-Power Industry 0.28 0.52 0.38 0.04 1.00 0.46 0.08 -0.15
South America-Processes 0.37 0.70 0.48 0.27 0.46 1.00 0.27 -0.18
South America-Transport 0.34 0.06 0.10 -0.00 0.08 0.27 1.00 0.10
South America-Waste -0.28 -0.21 -0.34 -0.21 -0.15 -0.18 0.10 1.00
Continent_Name-Sector

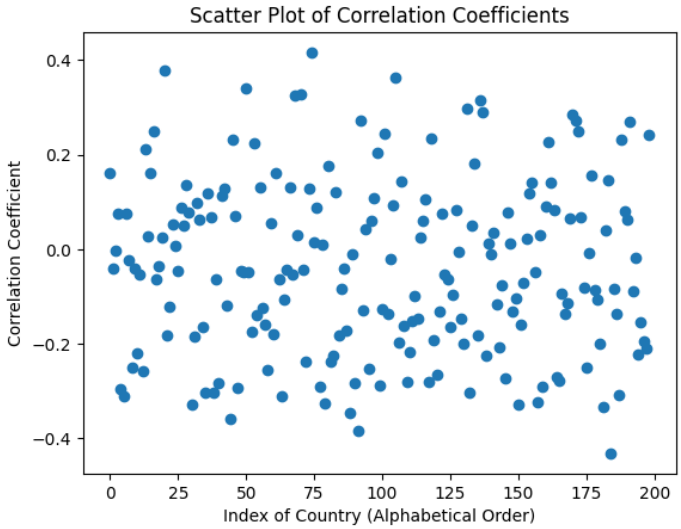
Emissions vs GDP Correlation

CO2 Emissions vs GDP

	Correlation
ABW	0.162023
AFG	-0.041652
AGO	-0.001594
AIA	0.074558
ALB	-0.296689
...	...
YEM	-0.222692
ZAF	-0.153741
ZMB	-0.195202
ZWE	-0.208954
EU27	0.242851

199 rows × 1 columns

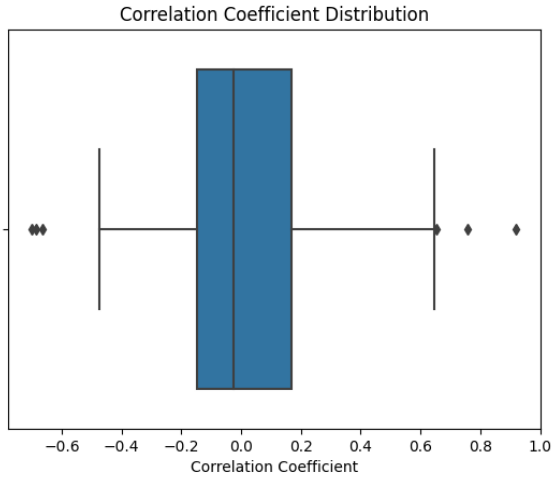


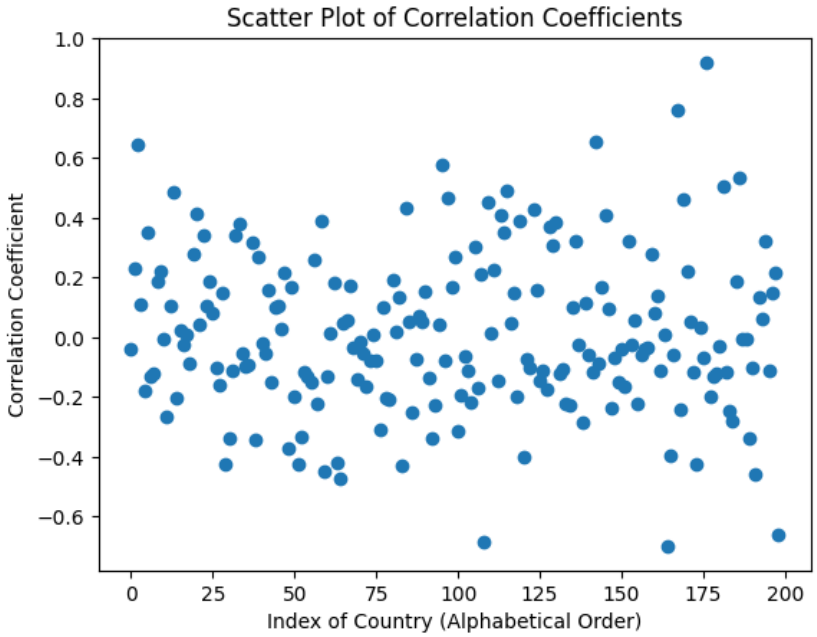


Greenhouse Gas Emissions vs GDP

	Correlation
ABW	-0.039314
AFG	0.231595
AGO	0.644282
AIA	0.109944
ALB	-0.180278
...	...
YEM	0.321663
ZAF	-0.110523
ZMB	0.147617
ZWE	0.217083
EU27	-0.662792

199 rows × 1 columns





To identify outliers, we used 2 standard deviations from the mean as a base. Should any data point be above this threshold, we will flag them out as outliers. This was done by passing in the stats.zscore method on the correlations .

	index	Correlation
2	AGO	0.644282
95	KGZ	0.578015
108	LTU	-0.685599
142	PNG	0.653077
164	SVK	-0.698678
167	SWZ	0.758982
176	TLS	0.919969
198	EU27	-0.662792

- Angola
- Kyrgyzstan
- Lithuania
- Papau New Guinea
- Slovakia
- Eswantini
- Timor-Leste

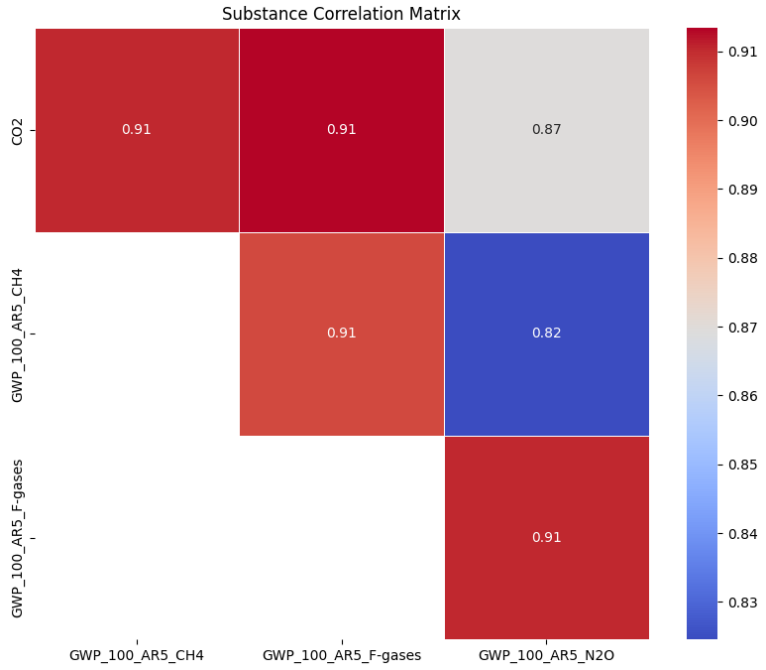
** EU27 refers to the European Union and was thus not classified as a country.

These anomalies all highlight a more positive or negative correlation which suggests both direct and inverse relationships respectively - positive correlation which suggests that the higher the GDP, the greater the level of GHG emissions, vice versa, negative correlation which suggests that the higher the GDP, the lower the level of GHG emissions.

- Timor Leste shows a visibly high correlation of 0.920 which could be possibly explained by the economy's dependence on oil and gas revenues. According to the China Dialogue in 2022, the Bayu Undan field in the Timor Sea finances about 85% of the government spending and the government is looking to develop the untapped reserves in the future. As a result, this cyclical dependence could have resulted in this high correlation.
- On the contrary, Slovakia shows the most negative correlation which is aligned with the general observation and policies within the country. According to Eurostat, there has been a "clear decoupling of GHG emissions from economic performance". This is due to the government's efforts to increase its renewable energy share and sectoral policies to enhance 'the carbon sink function' of the land-use change and forestry sector.

Emissions by Substances Correlation

Average Greenhouse Gas Emissions by Substances



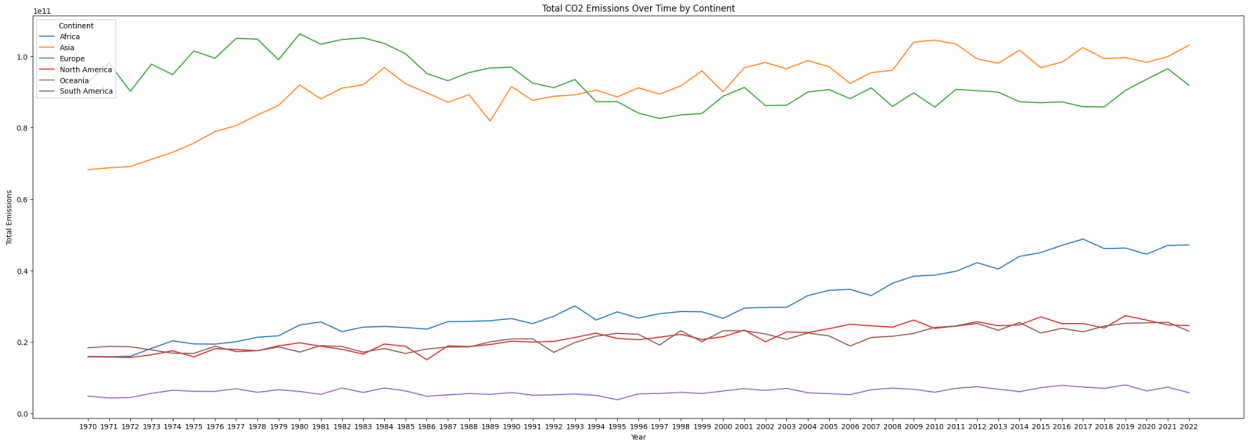
This correlation attempts to unravel the relationships and provide insights into how emissions of different substances interact. A positive correlation means that the higher the emission of substance A, the greater the emission of substance B as well. These can possibly indicate shared sources or similar factors influencing the emissions of both substances.

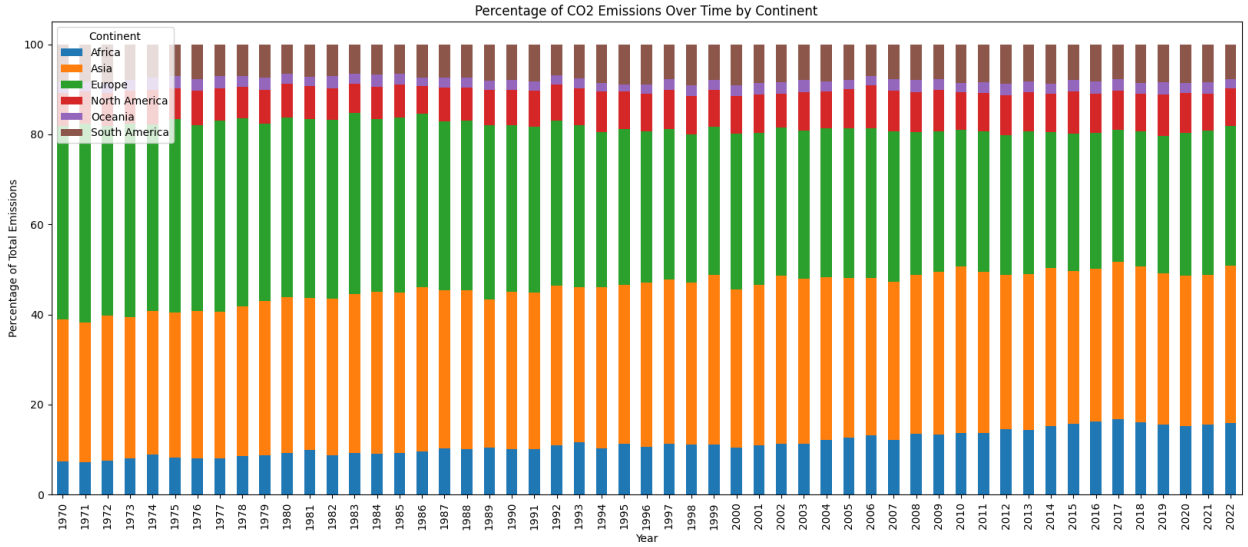
Based on the heatmap, we note a high positive correlation across the substances but most notably for CO2 and GWP_100_AR5_CH4 (methane) /_F-gases (fluorinated gases) at 0.91 and between the GWP substances. Interestingly, Nitrogen monoxide does not share a strong correlation with the other substances as expected.

As such, the next correlation will be to group by sectors and identify the common sources of emissions for the substances. This will provide a more granular perspective on the emission sources contributing to the releases of the substances.

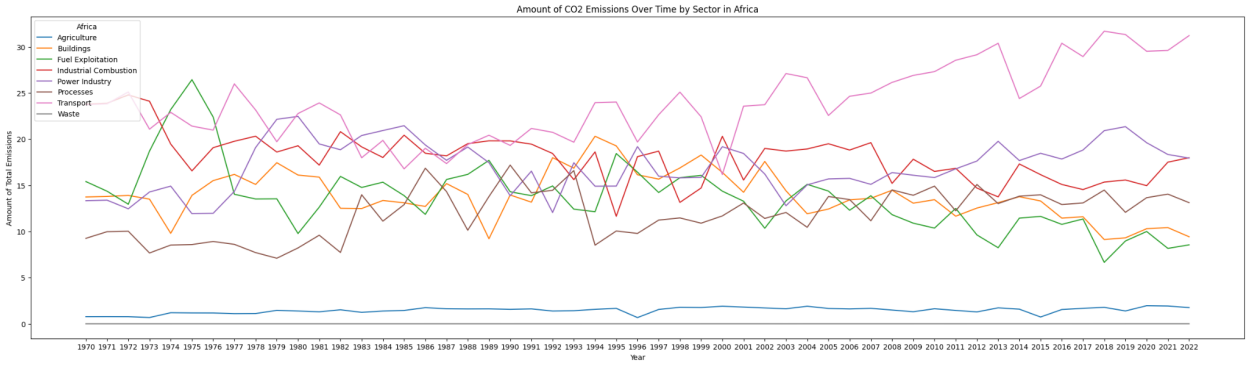
Temporal Analysis

CO2 Emissions

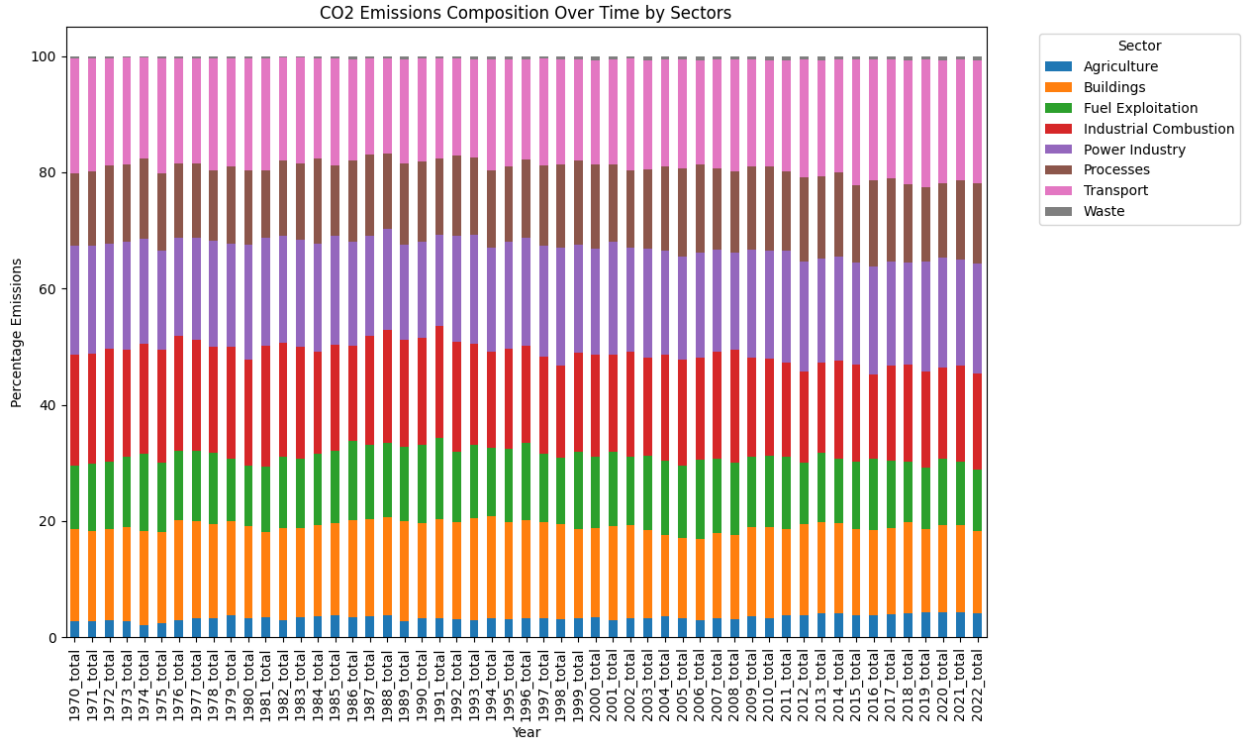




Over time, all continents recorded a rise in CO2 emissions with the exception of Europe. In the meanwhile, Africa led the pack with the greatest surge in CO2 emissions, nearly tripling throughout the time period. This can be possibly explained by the industrialization of Africa as the amount of CO2 emissions from transport rose dramatically in the continent. These include railways, buses and ferries as the countries experience urbanization and population growth, increasing the demand for transportation services. These signs of economic development often go in tandem with trade and business activities, necessitating transportation services to move goods and people. Furthermore, as Africa's GDP gradually increases, this enhances the affordability of citizens in the region to purchase vehicles, leading to elevated emissions. However, with the lack of government policies and public awareness, it fails to sufficiently curb these emission activities and promote sustainable emissions.

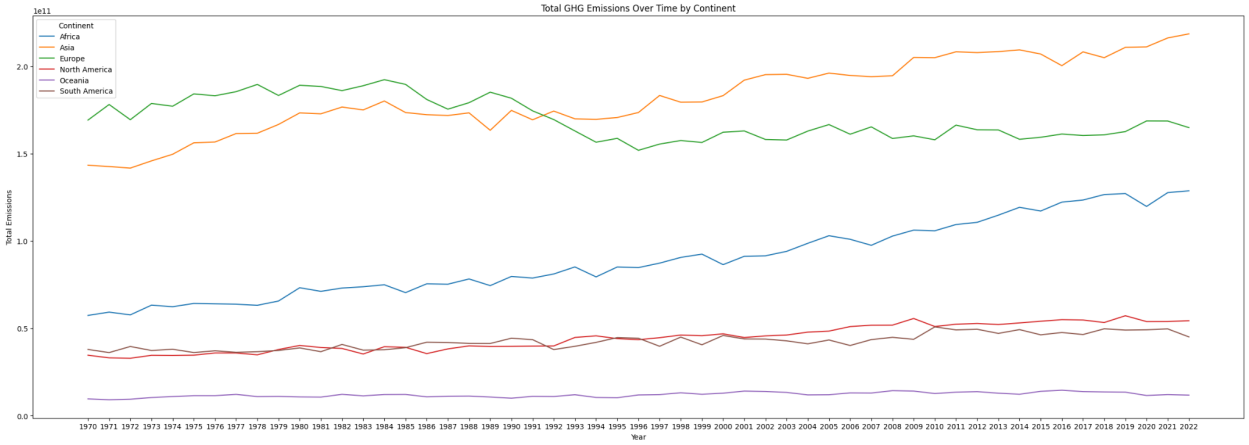


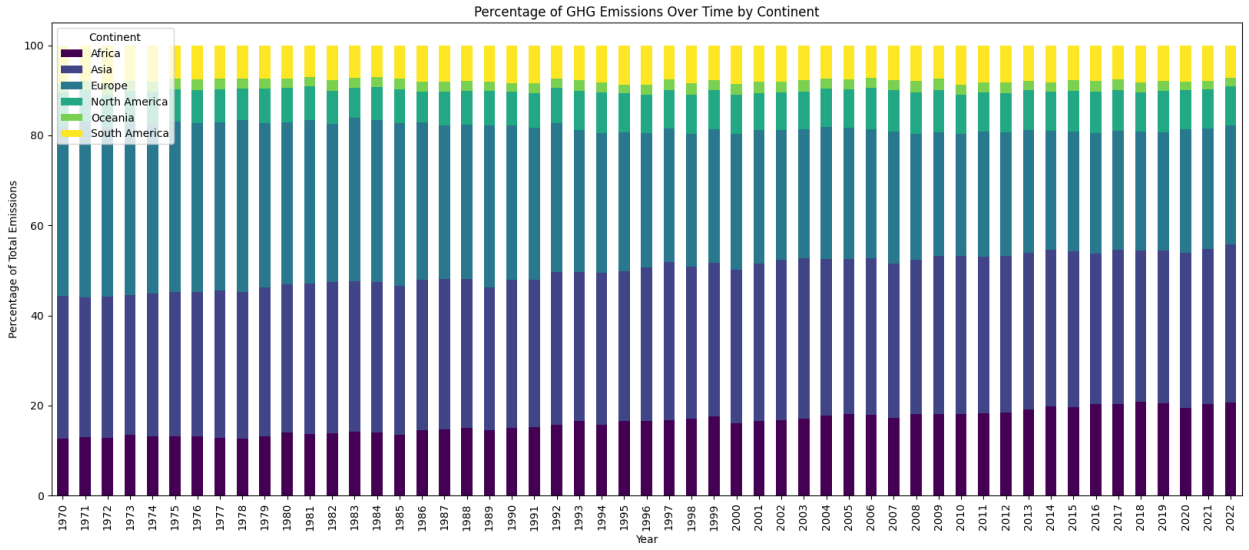
As for Europe which recorded a decrease in CO2 emissions, due to renewable energy investments, active government regulations and a carbon pricing market. Many European countries have made significant investments in renewable energy sources such as wind, solar, and hydropower. The increased share of clean and renewable energy in the energy mix helps reduce carbon emissions. In addition, strict environmental regulations and policies at the national and European Union levels have played a crucial role in reducing emissions. These policies include emission standards for industries, renewable energy targets, and commitments to international climate agreements. Finally, the European Union Emissions Trading System (EU ETS) is one of the world's largest carbon markets. It sets a cap on total emissions and allows companies to buy and sell emission allowances. This market-based approach provides economic incentives for companies to reduce their emissions and is a cornerstone of the EU's policy to combat greenhouse gas emissions cost-effectively.



On a global scale, the amount of emissions contributed by industrial combustion, fuel exploitation and buildings have decreased which are likely a result of the paradigm shift to embrace energy efficiency and alternative sources of energy. Technological advancements, such as the use of more efficient and cleaner technologies in industrial processes, contribute to emissions reduction. For example, the adoption of combined heat and power (CHP) systems and advanced manufacturing processes can improve energy efficiency and decrease emissions. Elsewhere, Green building standards and certifications, such as LEED (Leadership in Energy and Environmental Design), promote the construction of energy-efficient buildings. These buildings often incorporate features like better insulation, energy-efficient HVAC systems, and renewable energy installations, reducing emissions associated with heating, cooling, and electricity use.

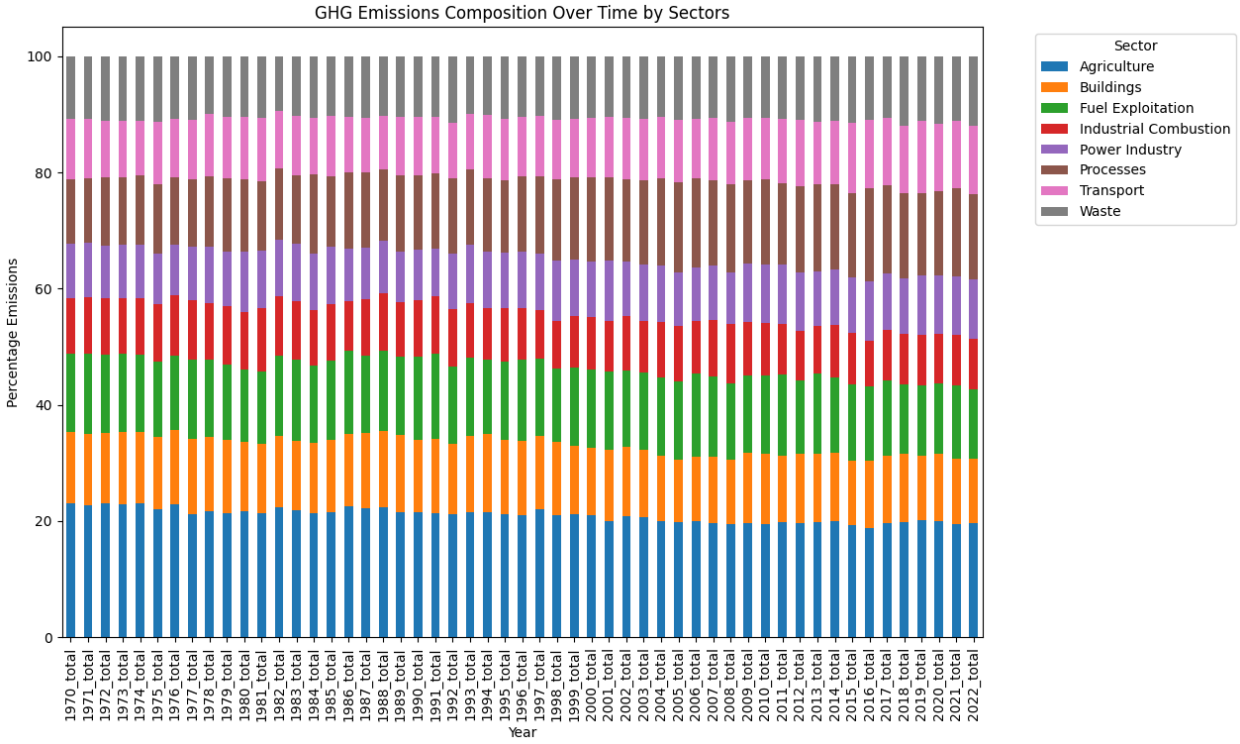
Greenhouse Gas Emissions





The same phenomenon is reflected for GHG emissions with Europe improving in its environmental efforts while the rest of the world continues to increase its emissions as Africa

leads the rise. In fact, Europe contributed 26.44% of total GHG emissions in 2022, down from 37.40% in 1970. On the other hand, Africa's contribution rose from 12.70% in 1970 to 20.65%. This dichotomy is largely reflected and expected based on the aforementioned analysis of CO2 emissions.



As for sectoral composition, agriculture led the greatest fall (23.10% to 19.59%) in contributions to total GHG emissions while Processes contributed the largest rise (10.94% to 14.64%). Livestock, especially cattle, contribute significantly to methane emissions. The implementation of improved livestock management practices, such as dietary changes and methane capture technologies, could contribute to lower emissions from the agriculture sector. As for processes, this could be plausibly explained by the changes in consumer demand and trade patterns which shifted to more energy intensive manufacturing techniques such as computer parts and electronic equipment, due to the rise in income per capita.

Prediction Model

Using the public dataset (`https://datahub.io/JohnSnowLabs/country-and-continent-codes-list/r/country-and-continent-codes-list-csv.csv`) which maps EDGAR country codes to the respective continent, we merged the dataframes to retrieve the continent for each country and determined the mean aggregated by continent on a yearly basis for the metrics - Africa, Asia, Europe, North America, Oceania and South America.

These functions have been packaged into a general function and the model can be easily generalized across continents for testing using both methodologies - Linear Regression and XGBoost algorithms. To test the accuracy of our model, for the XGBoost methodology, we introduced RMSE and MAPE errors to determine the error rate; for both methodologies, we have compared the predicted values against the actual values to calculate the percentage difference. These were charted alongside the scatter plots for each continent in the tables below:

Methodology: Linear Regression

Data Cleaning / Preparation

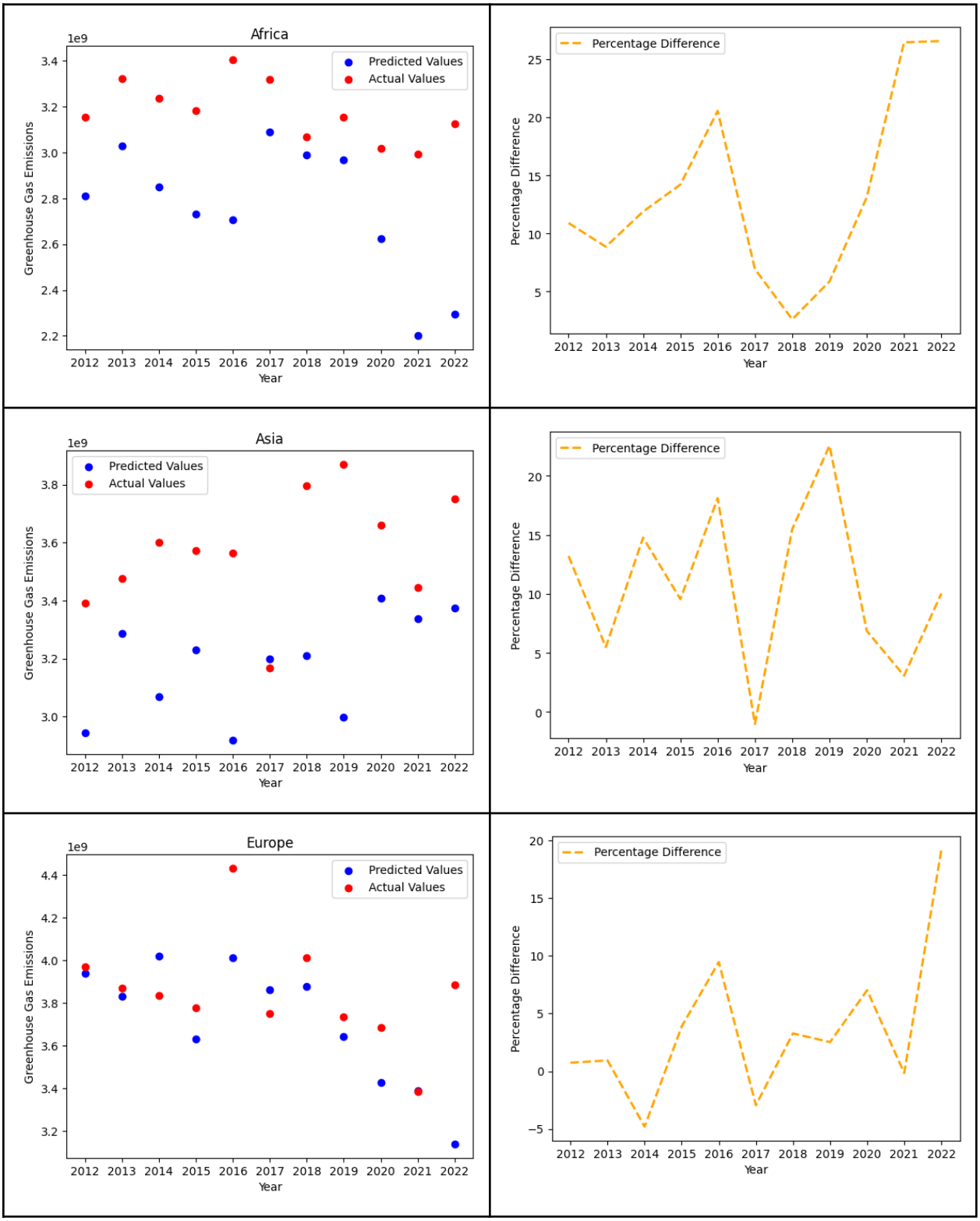
- The dataset was split into the following: Training Data (80%),and Test Data (20%). These different datasets were split and ordered by time to avoid the look-ahead bias which could lead to inaccurate results
- Finally, I performed feature scaling through the MinMax Scalar. This involves rescaling and shrinking the data of the features into a given range.

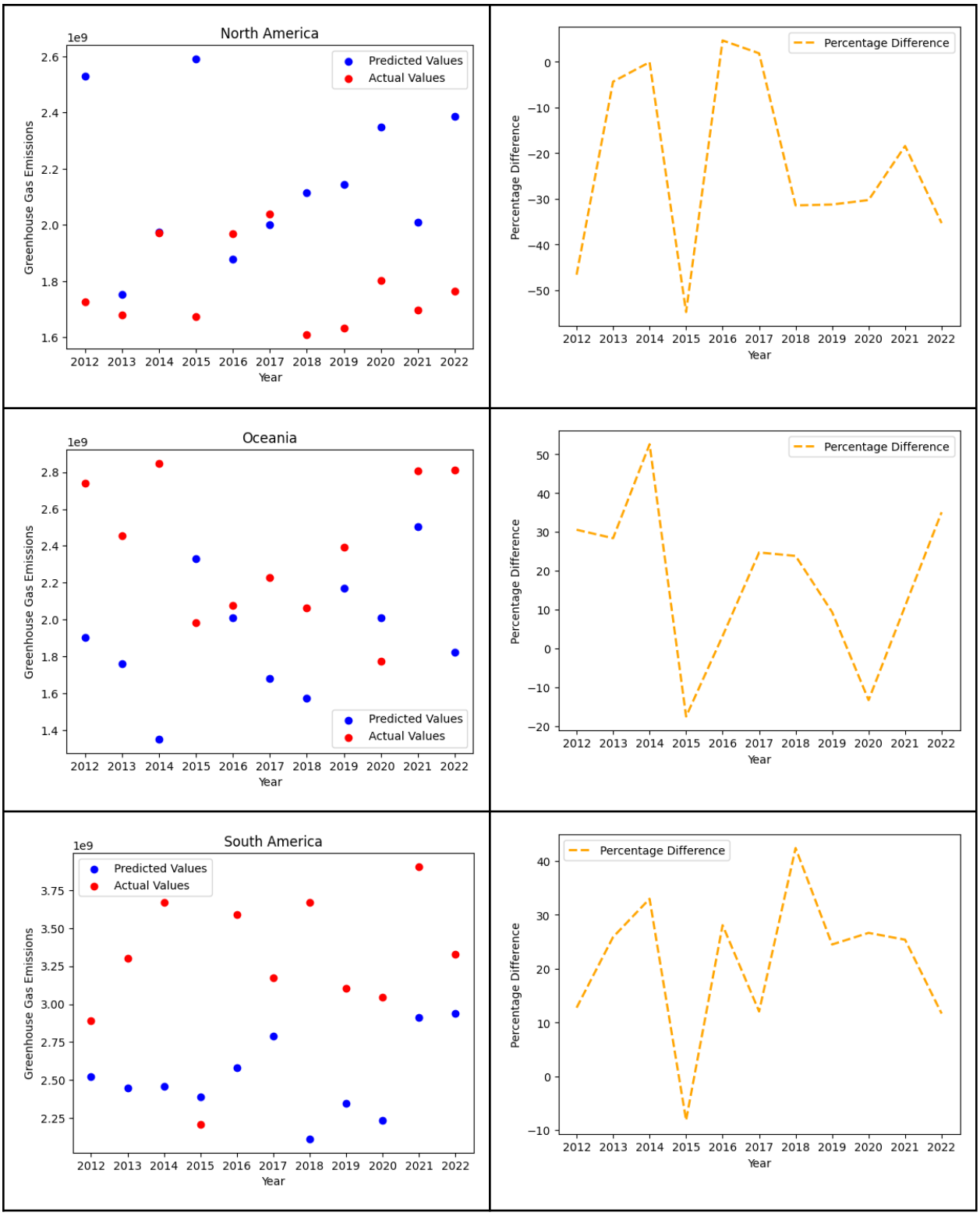
Linear Regression

Linear regression is used to predict the value of an independent variable based on a series of dependent variables. It estimates the coefficients of the linear equations and fits a straight line that minimizes the discrepancies between predicted and actual output values using the least squares method.

Results (Test Data Prediction)

GHG Emissions

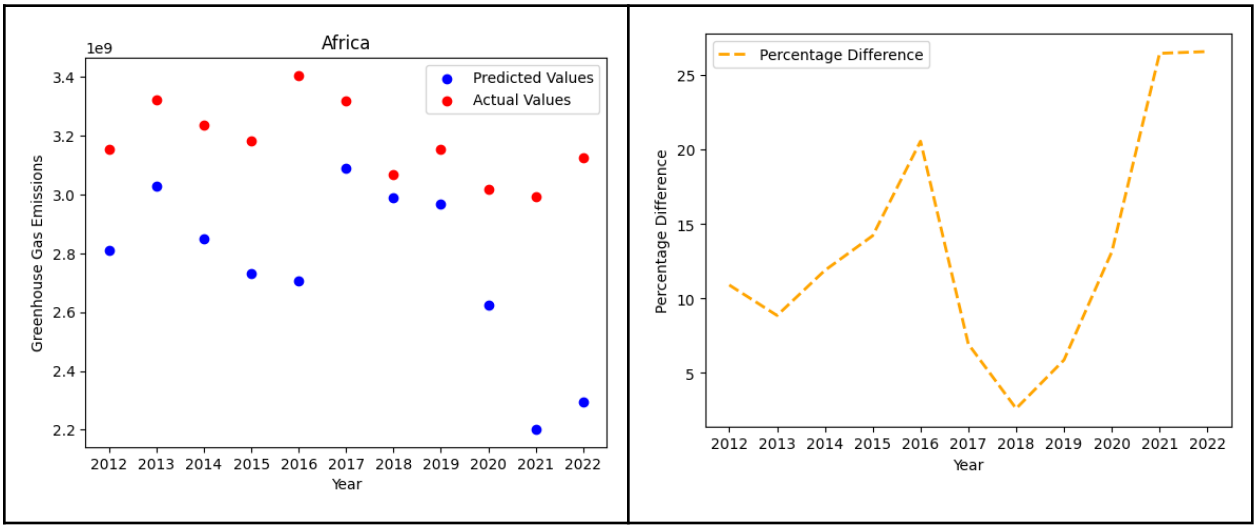

Africa
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year
Percentage Difference
Percentage Difference
Year
Asia
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year
Percentage Difference
Percentage Difference
Year
Europe
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year
Percentage Difference
Percentage Difference
Year

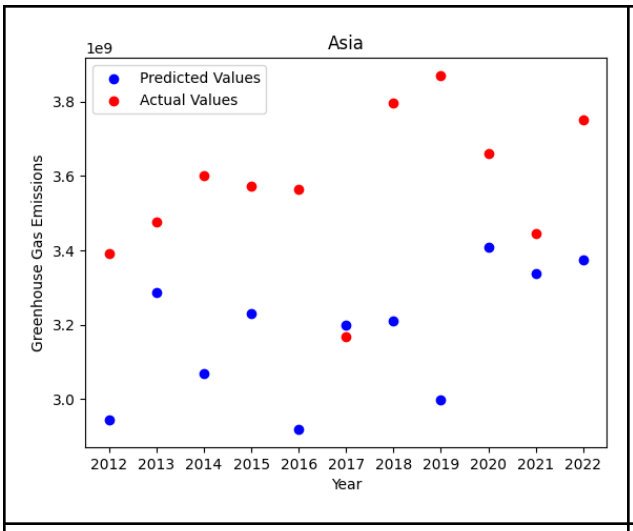


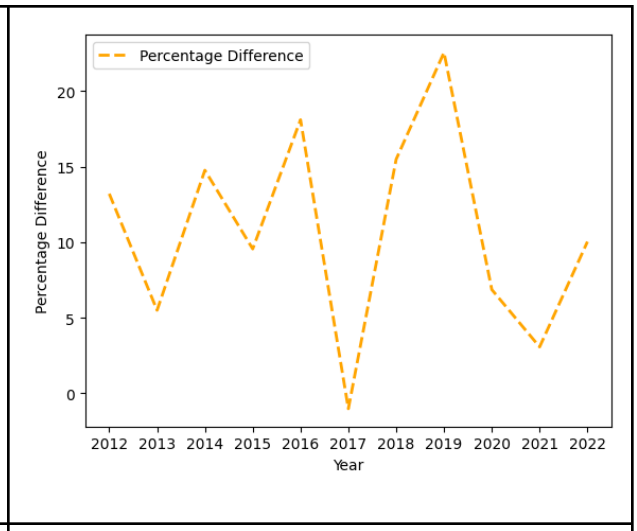
Regression Model Parameters

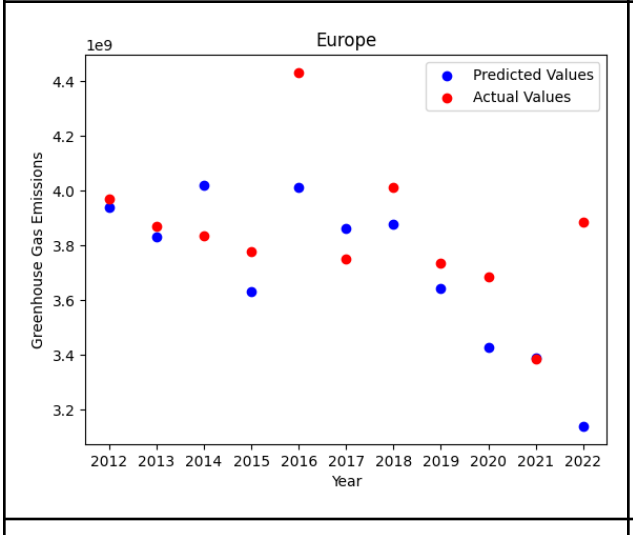
Feature	Africa	Asia	Europe	North America	Oceania	South America
capita	-285684164.79	-281975.25	-228303235.28	-265506245.84	151229982.66	-164372534.73
Agriculture	57835884 7.18	35000184.19	-409170571.81	24066953 8.58	-499058068.44	-100383601.49
Buildings	58049210 3.6	-145963057.38	-275014 39.3	-218029452.93	-585060044.49	285566564.96
Fuel Exploitation	-568568208.96	69910458.44	-124280536.02	-171935607.14	721110778.59	255931322.67
Industrial Combustion	-251141440.46	69281089.65	79723041.92	-208851584.42	234884891.86	-231462364.39
Power Industry	296542632.02	153218997.29	-127192316.48	-197325558.25	403401659.13	-358835908.77
Processes	-843570682.83	-8171006.15	-199540 51.02	114536099.51	-513848 06.26	-347045467.95
Transport	379434897.3	101248324.28	-447245068.01	388615251.58	248294853.58	421224575.1
Waste	327158509.81	360763371.09	67929575.44	-339528743.97	147484287.68	-954080577.33
gdp	-230346498.31	-278530 91.06	259417817.01	-659349591.61	386912200.79	479308023.16
Intercept	2780178329.34	2848026028.61	4111954173.42	2624652141.91	1372413307.07	3155586649.59

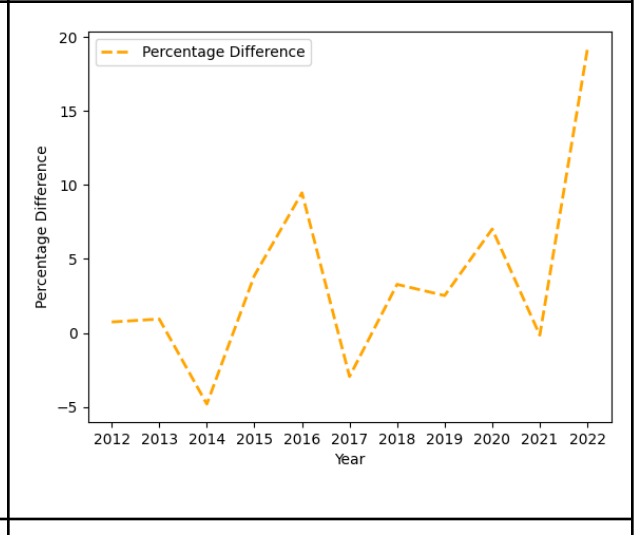
CO2 Emissions

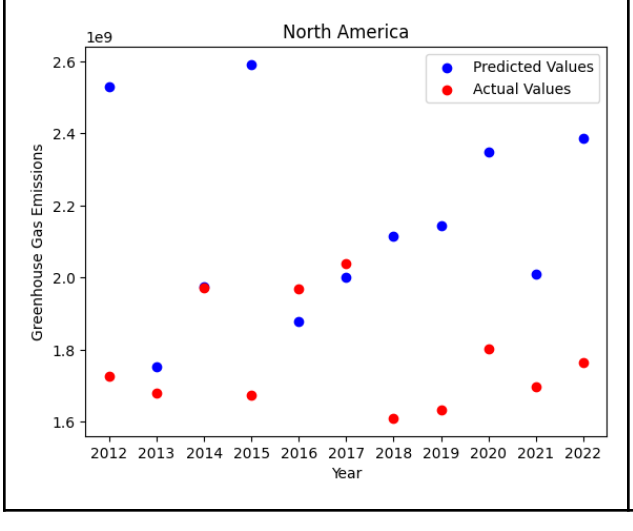


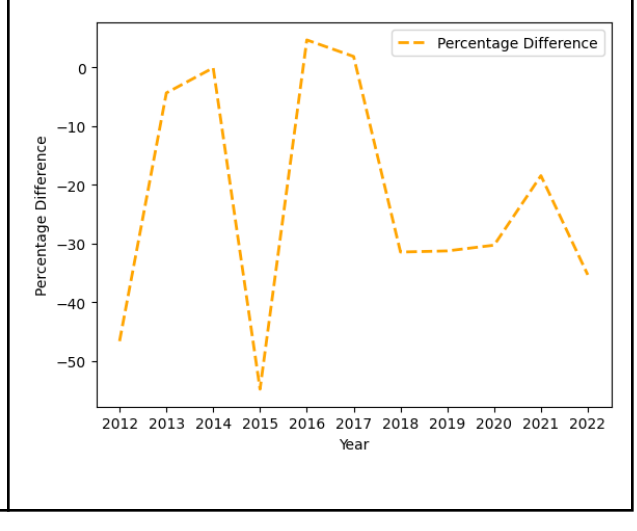

Asia
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year

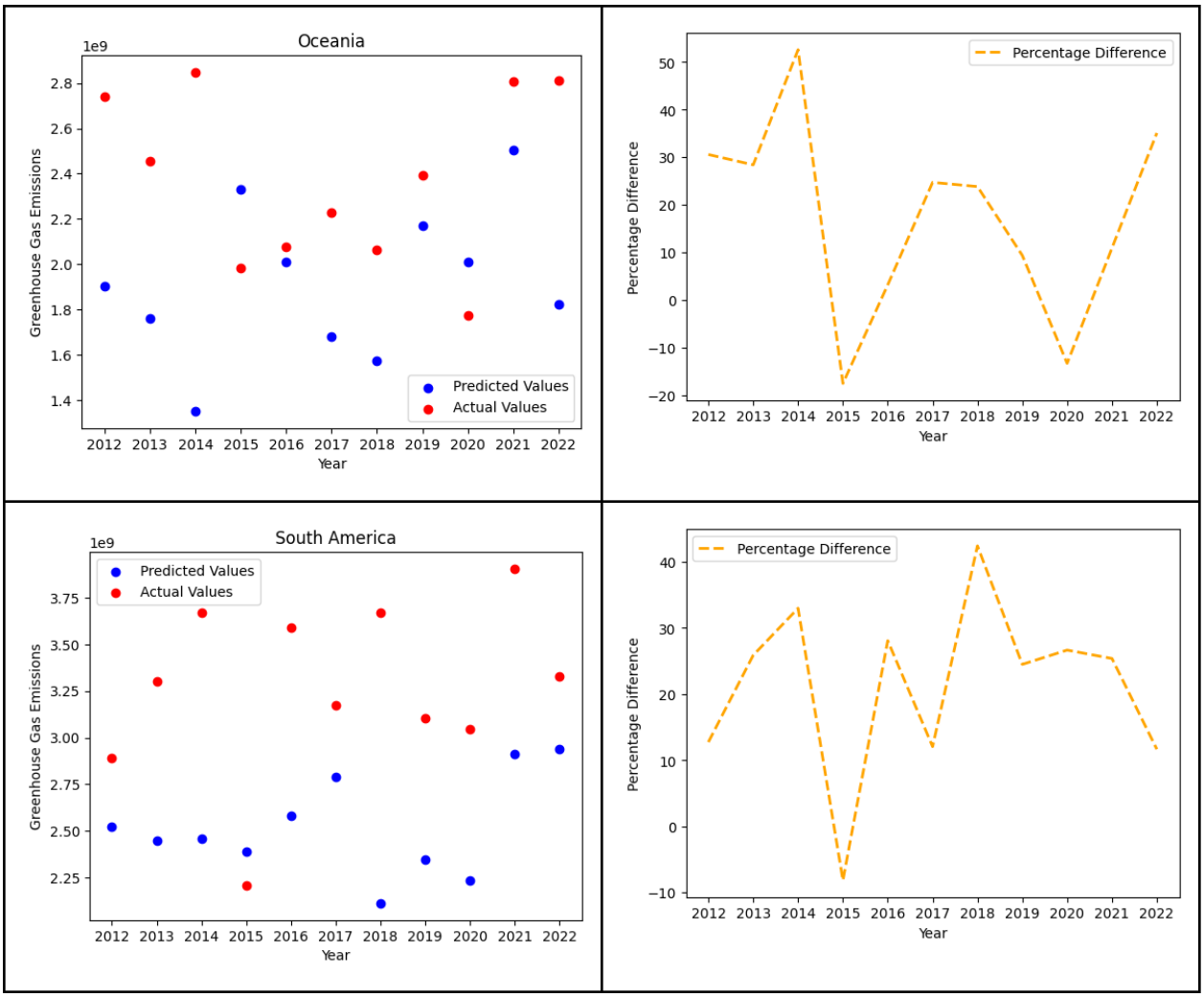

Percentage Difference
Percentage Difference
Year


Europe
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year


Percentage Difference
Percentage Difference
Year


North America
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year


Percentage Difference
Percentage Difference
Year



Regression Model Parameters

Feature	Africa	Asia	Europe	North America	Oceania	South America
capita	552256143.03	18559662.26	185321076.53	-332618291.68	135501413.81	732667571.99
Agriculture	-1489475.39	343676942.48	-295047586.83	-32293464.87	250965963.68	-910940803.57
Buildings	131291984.91	-358518749.61	-105703819.84	-619627 2.35	-298384420.66	-82576415.29
Fuel Exploitation	-168404138.2	163201096.79	-159839055.95	219391380.77	638909237.58	920461496.94
Industrial Combustion	-487817842.17	37689289.43	-72303543.98	-145849 00.75	200202422.57	-112160201.18
Power Industry	574999685.53	-473080400.84	-184068437.43	61419005.7	398547018.16	-457035922.55

Processes	-125135471.3	-1208300 7.96	21787623 1.17	34454069.98	-233296454.82	2629771 68.94
Transport	2611974 21.48	1317808 44.89	-3030011 83.34	1312935 60.81	2295654 90.44	-245719 263.11
Waste	-490918 258.5	2616383 82.61	28542560.49	-501503 19.84	55195887.01	-2258114 98.97
gdp	31086236.37	-100918 057.82	3331510 95.54	-776937 932.53	1712351 90.93	7337712 94.14
Intercept	2694334 229.85	2901946 994.3	3773566 807.14	2456397 692.46	1015939 255.15	2473851 984.58

Methodology: XG Boost

Data Cleaning / Preparation

- The dataset was split into the following: Training Data (60%), Validation Data (20%) and Test Data (20%). These different datasets were split and ordered by time to avoid the look-ahead bias which could lead to inaccurate results
- Finally, I performed feature scaling through the MinMax Scalar. This involves rescaling and shrinking the data of the features into a given range. It is useful for gradient descent algorithms since a huge difference in ranges of features will cause vastly different step sizes for each feature. Therefore, having features on a similar scale will help the gradient descent converge more quickly towards the minima. Since the distributions of the various features are not Gaussian, StandardScalar was thus omitted as an option.

XGBoost

XGBoost is a decision tree based ensemble model that adopts a gradient boosting framework. Presented by Chen and Guestrin in 2016, the algorithm mainly enhances existing boosting models through:

a. Regularisation : Penalises more complex models through L1 and L2 regularisation to prevent overfitting and smoothens the final learnt weights

b. Shrinkage Estimates : Scales newly added weights after each step of tree boosting through column sub-sampling. It reduces the influence of each individual tree and leaves space for future trees to improve the model

Using the scaled dataset, I iteratively applied the XGBRegressor on the scaled training dataset to train the model and tune the parameters. This was then predicted on the scaled dataset and scaled back to the original range to obtain the total greenhouse gas emissions. To

choose the optimum value for each parameter at each iteration, I picked the value which corresponds with the minimum RMSE.

Iteration 1 : Default Parameters

I initiated the model using an XGBRegressor with default values of each variable and the objective function as ‘reg:squared error’ - regression with squared loss.

Iteration 2 : n_estimators, max_depth

- n_estimators : Range of 10 to 100, with a step of 5
- max_depth : Range of 1 to 10, with a step of 1

Iteration 3 : learning_rate, min_child_weight

- learning_rate : Range of 0.01 to 1 with a step of 0.01
- min_child_weight : Range of 1 to 21 with a step of 1

Iteration 4 : subsample, gamma

- subsample : range of 0.1 to 1 with a step of 0.1
- gamma : range of 0.1 to 1 with a step of 0.1

Iteration 5 : colsample_bytree, colsample_bylevel

- colsample_bytree : range of 0.5 to 1 with a step of 0.1
- colsample_bylevel : range of 0.5 to 1 with a step of 0.1

To evaluate the performance of the model, the following metrics were used:

Accuracy

a. Mean Absolute Percentage Error (MAPE)

In the time series forecasting model, I used MAPE to determine the absolute error and avoid any model or parameters with a high absolute error. However, MAPE would not work well alone as it cannot detect the error of zeros and extreme values. Hence, I need to use RMSE along with MAPE.

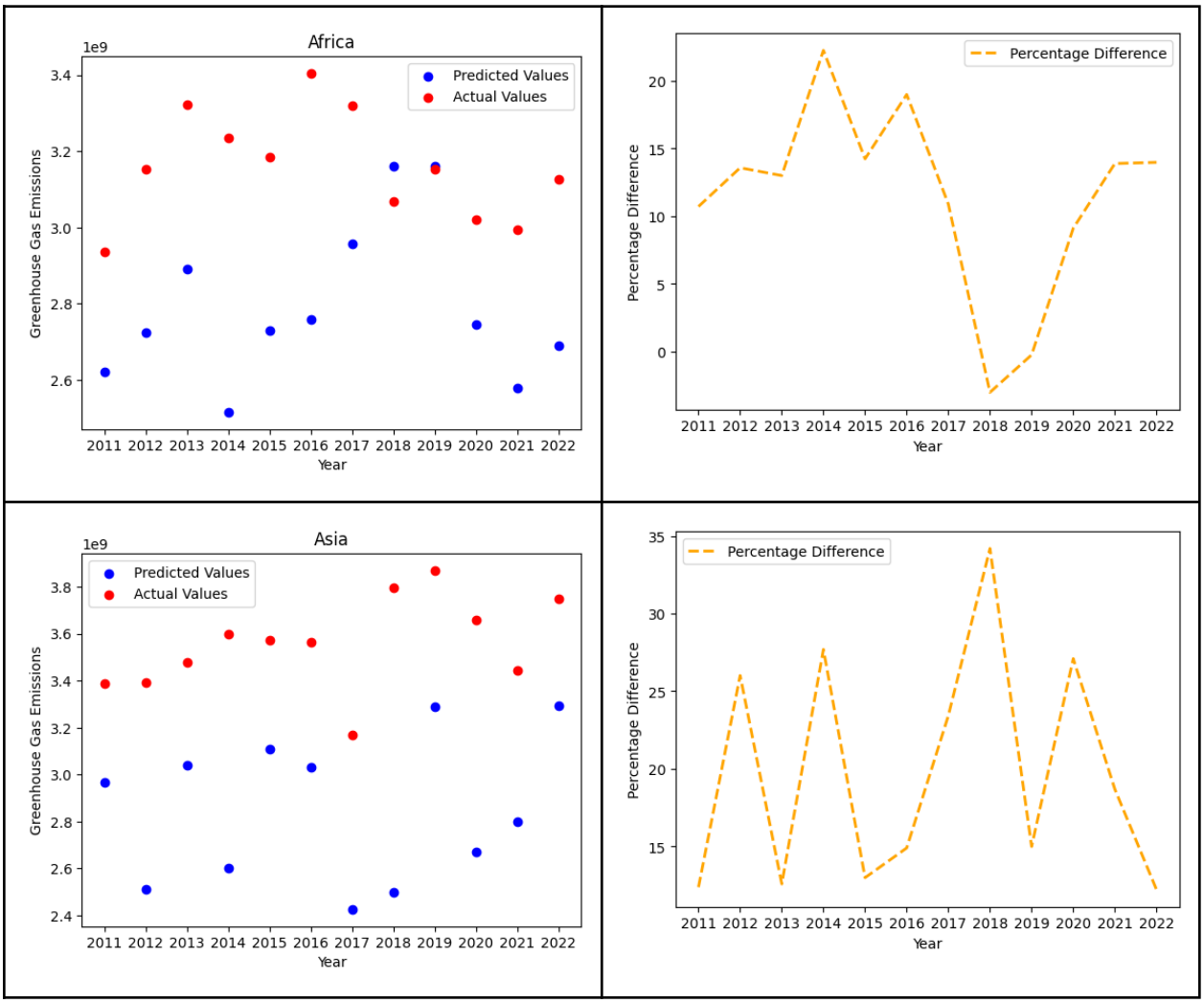
b. Root Mean Square Error (RMSE)

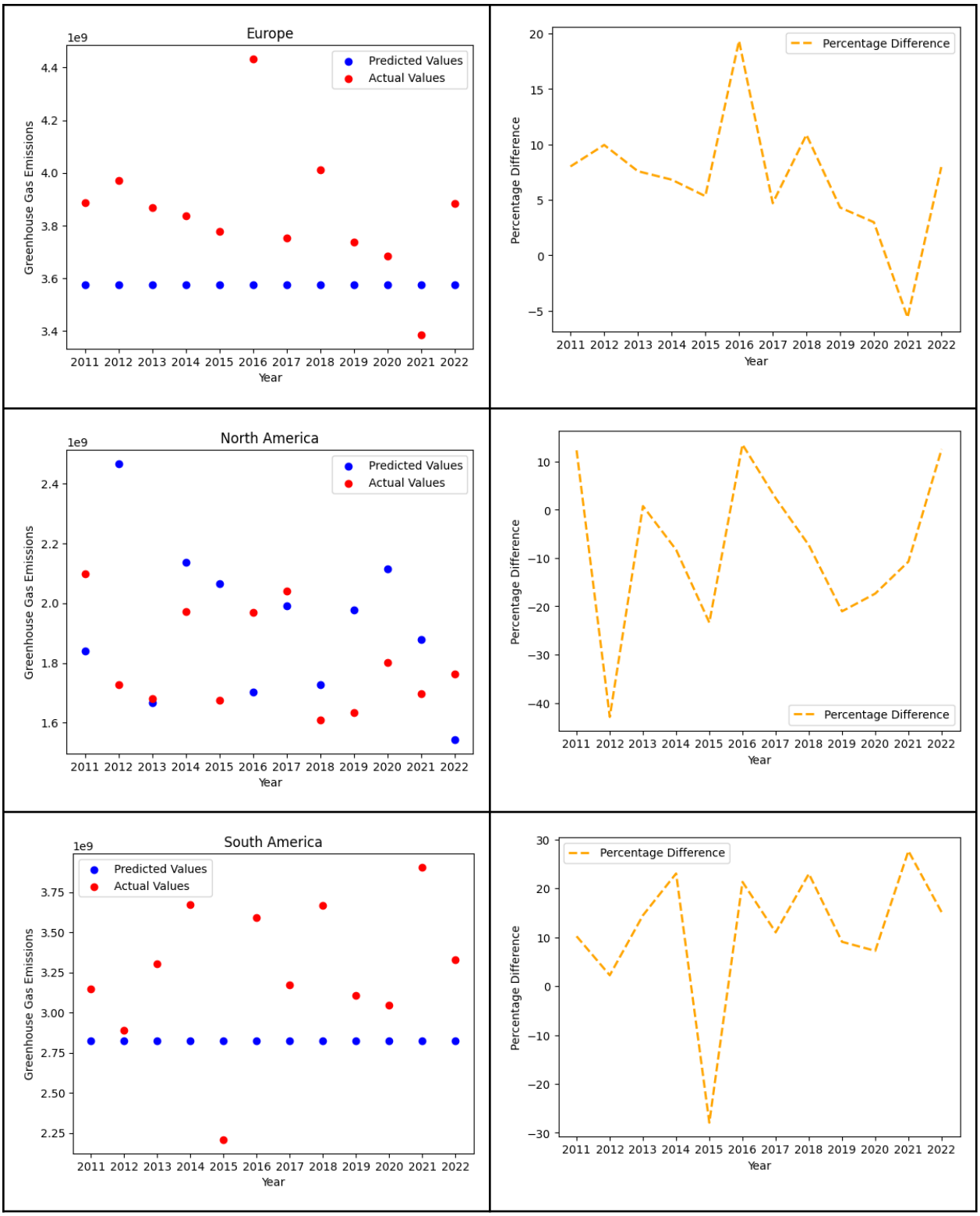
RMSE measures the standard deviation between the predicted values from the model and the actual values of a dataset. The higher the standard deviation, the larger the error. Similar to MAPE, RMSE will always be positive and a lower value indicates higher performance.

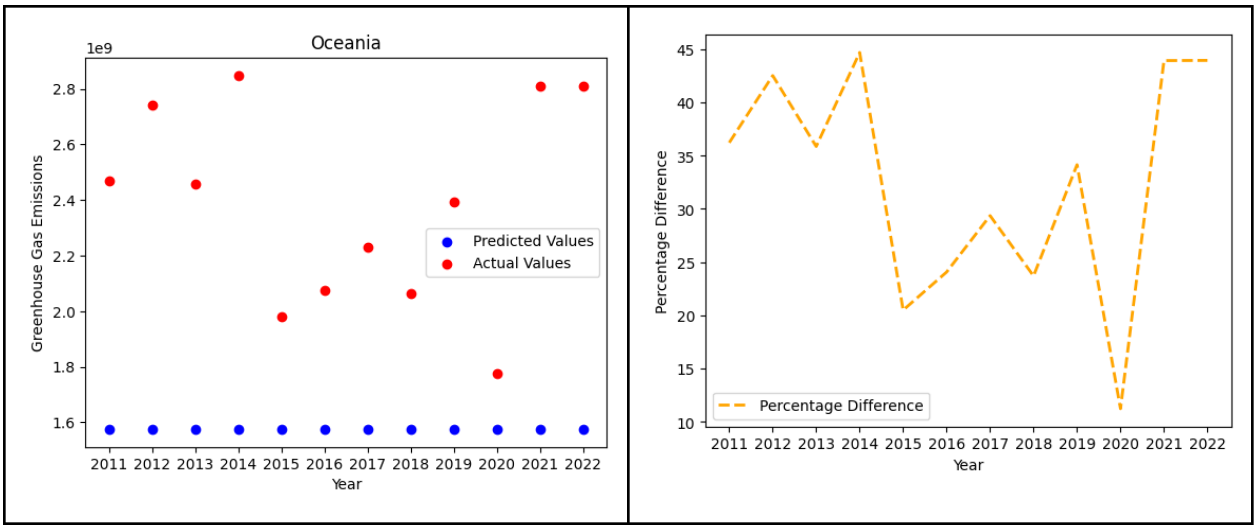
**For all the hyperparameters tuning, only RMSE was used as the minimum. While both MAPE and RMSE summarize the variability of the observations around the mean, they are of different scale, hence contributing to the vastly different values. RMSE was therefore chosen since it is the basis for how the model fits the data.

Results (Test Data Prediction)

GHG Emissions



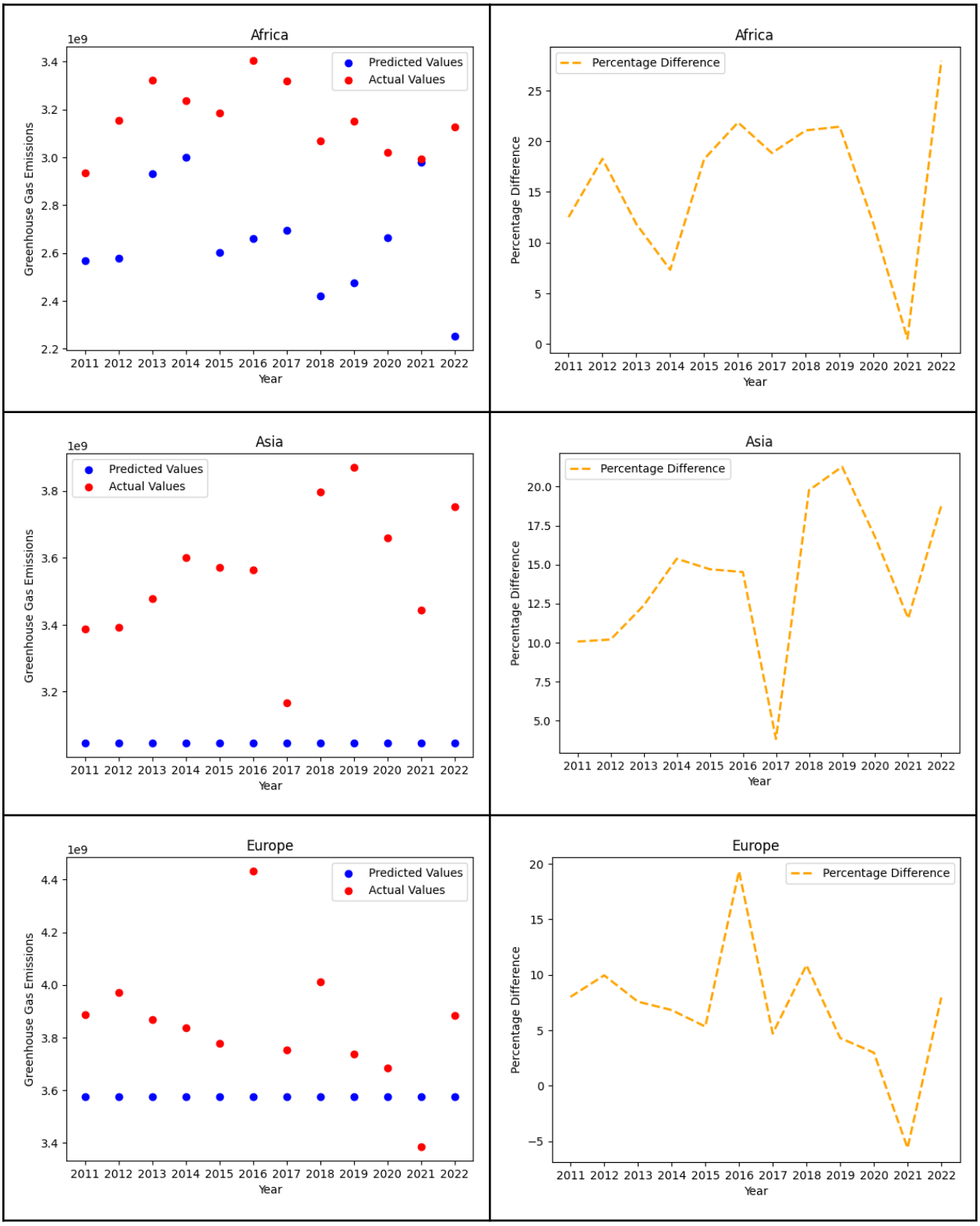

Europe
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year
Percentage Difference
Percentage Difference
Year
North America
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year
Percentage Difference
Percentage Difference
Year
South America
1e9
Predicted Values
Actual Values
Greenhouse Gas Emissions
Year
Percentage Difference
Percentage Difference
Year

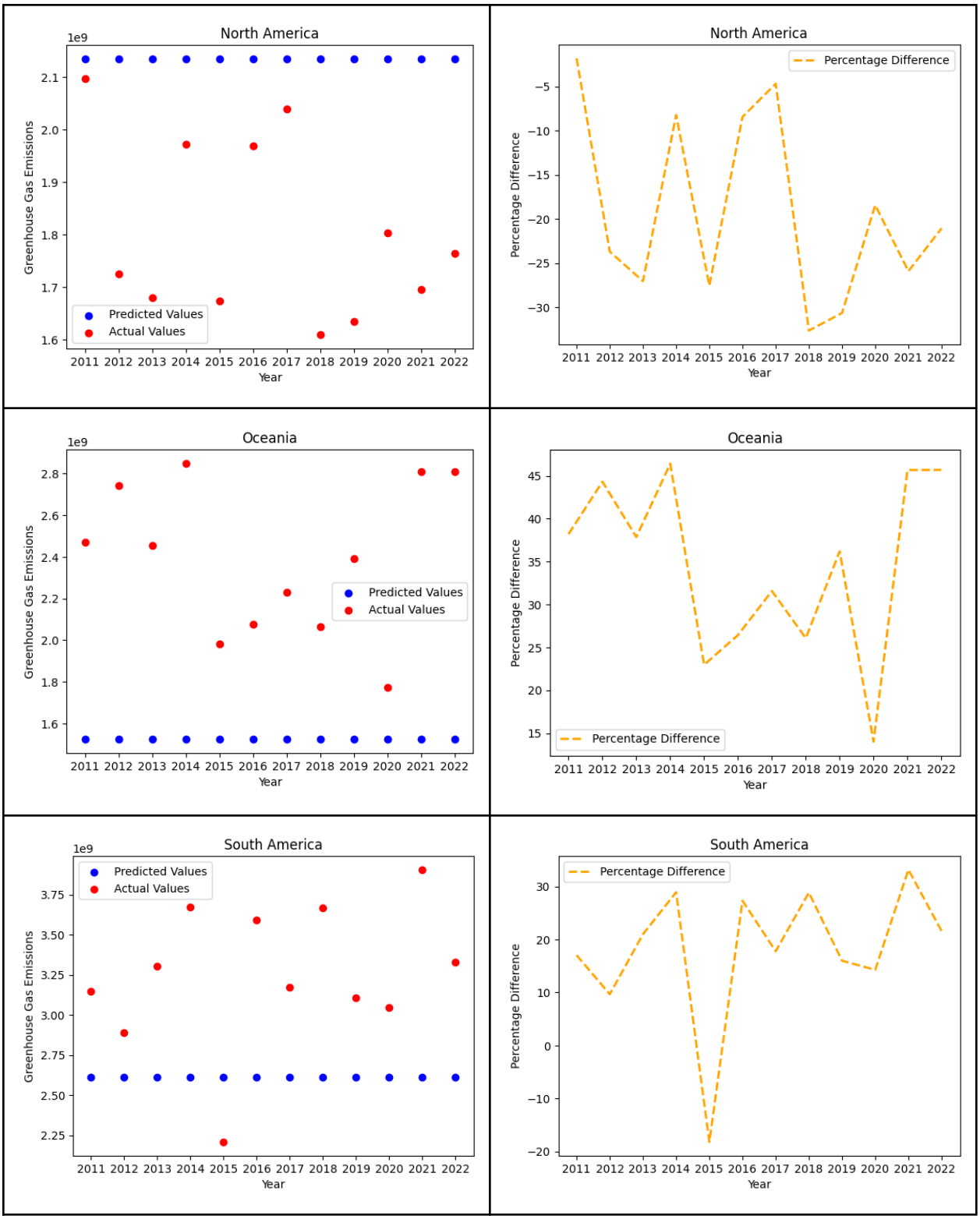


Optimal Parameters

Parameters	Africa	Asia	Europe	North America	Oceania	South America
n_estimators	30.0	40.0	10.0	95.0	10.0	35.0
max_depth	6.0	7.0	3.0	2.0	1.0	8.0
learning_rate	0.87	0.94	0.01	0.99	0.01	0.88
min_child_weight	10.0	4.0	17.0	11.0	17.0	14.0
subsample	0.6	0.5	0.6	0.9	0.6	0.1
gamma	0.6	1.0	0.5	0.9	0.5	0.5
colsample_bytree	0.5	1.0	0.5	0.6	0.5	0.5
colsample_bylevel	0.01	0.01	0.01	0.01	0.01	0.01

CO2 Emissions




North America
1e9
Greenhouse Gas Emissions
Predicted Values
Actual Values
Year
North America
Percentage Difference
Percentage Difference
Year
Oceania
1e9
Greenhouse Gas Emissions
Predicted Values
Actual Values
Year
Oceania
Percentage Difference
Percentage Difference
Year
South America
1e9
Greenhouse Gas Emissions
Predicted Values
Actual Values
Year
South America
Percentage Difference
Percentage Difference
Year

XGBoost Model Optimal Parameters

Parameters	Africa	Asia	Europe	North America	Oceania	South America
n_estimators	20.0	10.0	10.0	55.0	10.0	15.0
max_depth	6.0	1.0	3.0	8.0	4.0	2.0
learning_rate	0.95	0.01	0.01	0.72	0.48	0.99
min_child_weight	1.0	13.0	17.0	9.0	16.0	9.0
subsample	0.4	0.1	0.6	0.3	0.5	0.2
gamma	1.0	0.5	0.5	0.5	0.5	0.5
colsample_bytree	0.7	0.5	0.5	0.5	0.5	0.5
colsample_bylevel	0.01	0.01	0.01	0.01	0.01	0.01