Team 7: Erica Zhao, Harshavardana Reddy, Yash Manish Doshi, Sai Pavan

Subject: DATS 6103 Introduction to Data Mining

Professor: Sushovan Majhi

18 December 2023

A Comprehensive Analysis of Global Greenhouse Gas Dynamics

This report offers an exhaustive analysis of greenhouse gas (GHG) emissions from 2016 to 2021, focusing on sectors and gasses contributing significantly to global emissions. Using datasets from the IMF and UNFCCC, we analyzed annual country-level GHG emissions across various sectors. The study employs advanced statistical models, including Linear Regression and ARIMA, to dissect industry contributions, scrutinize emission trends in major economies, assess methane mitigation efforts, and explore the disparity in emissions between top and bottom emitting countries. Our findings, which reveal critical insights for policymakers and stakeholders, underscore the need for strategic intervention in climate policy development, particularly in high-emitting sectors and economies.

1. Introduction

In an era where climate change poses a paramount challenge, understanding the intricacies of greenhouse gas emissions is crucial. This report embarks on a detailed exploration of GHG emissions, dissecting their sources, trends, and potential future trajectories. Leveraging a rich dataset from the IMF and UNFCCC, this study aims to provide a comprehensive understanding of the global emissions landscape. The analysis covers annual emissions across various sectors and countries, emphasizing the years 2016 to 2021. Through this investigation,

we seek to offer actionable insights into the sectors and gasses that contribute most to GHG emissions and evaluate the effectiveness of existing mitigation strategies.

2. Smart questions

1. Which industries contribute the most to greenhouse gas (GHG) emissions, in the dataset. Which gas type is emitted most in every Industry?

2. What has been the pattern of GHG emissions from 2016 to 2021 for top 10 economy countries and which country has shown the greatest deviation from this trend?

3. Among all the countries examined in the dataset, which ones have achieved a reduction in methane emissions between 2016 and 2021 indicating efforts in mitigation?

4. In terms of emissions, how do the top 10% highest emitting countries, in 2021, compare to the bottom 10% and what factors could explain this discrepancy?

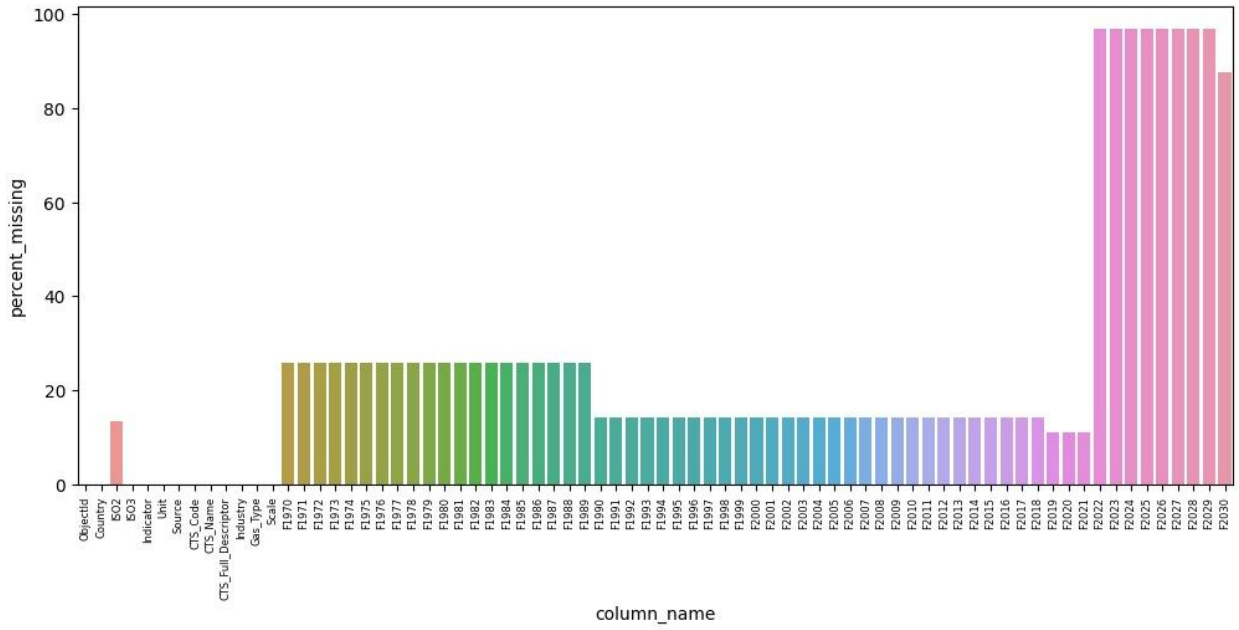
3. Data Description

Our study utilizes a comprehensive dataset provided by the United Nations Framework Convention on Climate Change (UNFCCC) and the International Monetary Fund (IMF), offering an extensive view of global greenhouse gas emissions. This dataset, encompassing records from 215 countries over the period 1970 to 2021, includes detailed data across six sectors and 14 sub-sectors. It comprehensively covers four primary greenhouse gasses, providing a nuanced perspective of global emissions. A unique aspect of this dataset is its presentation of emissions data both including and excluding land-use, land-use change, and forestry (LULUCF). This dual presentation allows for a deeper analysis of emission sources. Moreover, the dataset includes projections up to 2030 under a business-as-usual scenario, alongside national mitigation targets for the year 2030, offering insights into potential future emission trajectories. During our

analysis, we encountered challenges such as significant data gaps in future projections, which necessitated a focus on historical data for a more reliable and accurate study.

4. Data Preparation and Exploratory Data Analysis

In preparing the data for analysis, we meticulously cleaned the dataset, removing columns with projected data beyond 2021 and addressing the issue of missing values to ensure the integrity and reliability of our analysis. A critical step in our exploratory data analysis was visualizing the completeness of the dataset. We employed bar plots to highlight the percentage of missing values across different columns, revealing substantial data gaps in future projections (2022-2030). Consequently, we focused our analysis on the historically reliable data from 2016 to 2021. Additionally, we streamlined the dataset by discarding unrelated columns such as 'ISO2', 'CTS_Full_Descriptor', and 'CTS_Code', to focus on the most relevant variables. This phase of exploratory analysis was pivotal in understanding the distribution of emissions across different countries and industries. By grouping the data by country and industry category, we were able to identify the largest contributors to GHG emissions and the most emitted gas types in each sector, setting a solid foundation for our in-depth analysis of the emissions data.



Missing Data Analysis by Column in Dataset

- The bar graph shows the percentage of missing data for different columns in a dataset, indicating data completeness and the need for data cleaning or imputation strategies.
- Columns labeled from 'F1970' to 'F2021' have consistently low missing data, while 'F2022' to 'F2030' show a dramatic increase, suggesting possible issues with data collection or recording in these fields.

- Varying Levels of Missing Data: The columns exhibit a wide range of missing data percentages. Some columns have almost no missing data, while others have up to 100% missing.
- Columns with Minimal Missing Data: The column, 'ISO2', has very low percentages of missing data, indicated by the short bars at the beginning of the chart. This suggests that these fields are almost always populated.

- Columns with Moderate Missing Data: There is a group of columns, labeled from 'F1970' to 'F2021', which have a moderate amount of missing data, approximately 20-30%. This suggests that while these fields are often populated, there are still significant gaps in the data.
- Columns with High Missing Data: Towards the end of the chart, there are columns, particularly 'F2022' to 'F2030', which have very high percentages of missing data, indicated by the tall bars. Most of these fields lack data due to the upcoming years and the unavailability of the data yet.
- Pattern of Missing Data: There seems to be a pattern in the data where consecutive columns (likely representing similar types of data or collected in a similar fashion) have similar levels of missing data. This could be indicative of a systematic issue in how data is collected or reported for these fields.
- 100% Missing Data: The columns at the very end of the chart have bars reaching 100%, which means there is no available data in these columns at all. Due to the unavailability of data in the upcoming years, they are effectively empty.

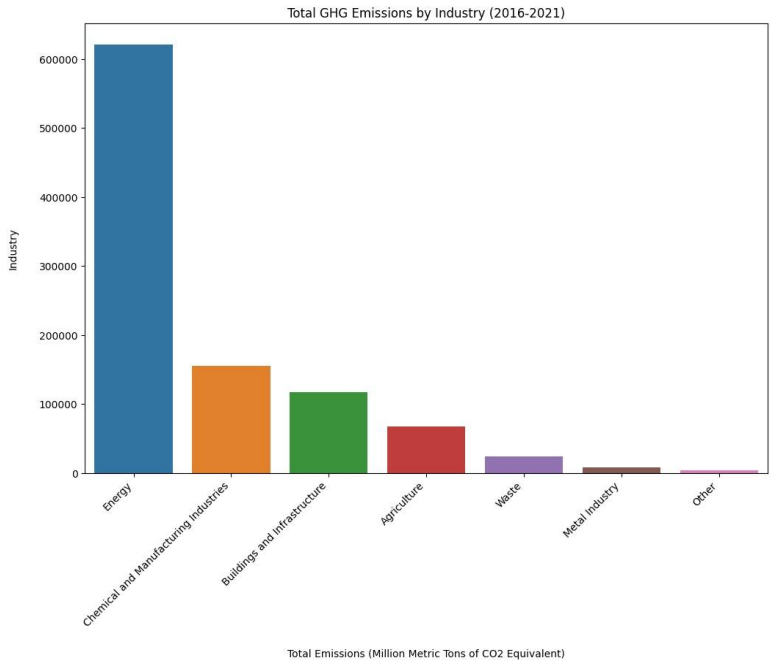
5. **Smart questions 1:**

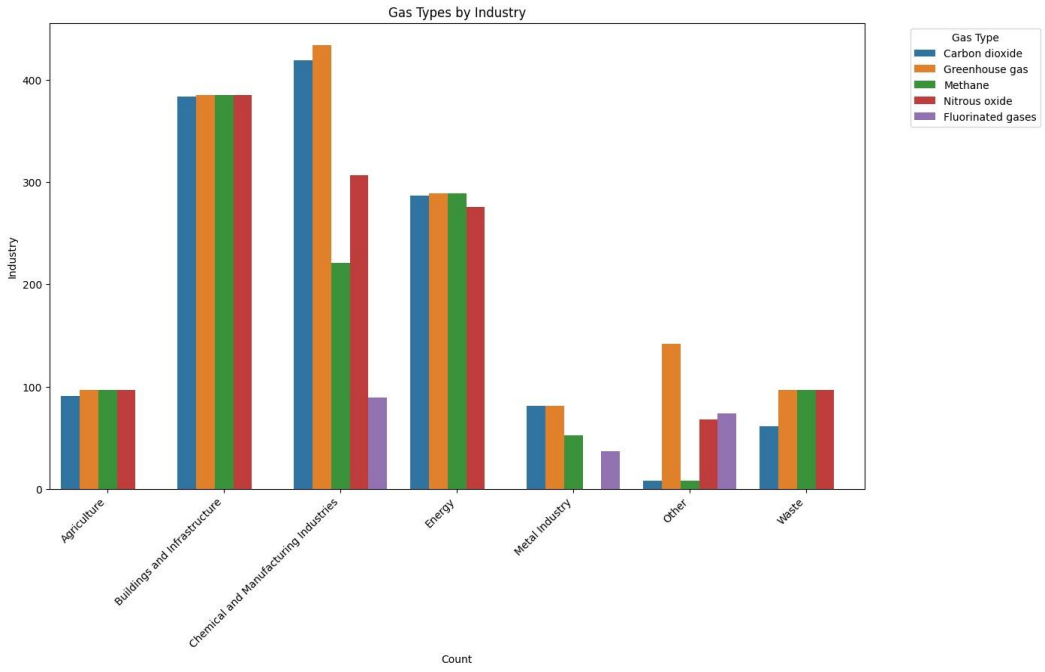
Which industries contribute the most to greenhouse gas (GHG) emissions, in the dataset. Which gas type is emitted most in every Industry?

Analysis by Industry and Gas Type

Our analysis commenced with a thorough examination of GHG emissions by industry. Utilizing the data analysis capabilities of Python, we grouped the data into industry categories derived from the detailed sub-sectors provided in the dataset. This categorization process

encompassed sectors such as Agriculture, Energy, Waste, and several others. We then aggregated the emissions data from 2016 to 2021, determining the overall contribution of each industry to GHG emissions. A bar plot illustrating these aggregates highlighted that the Energy sector emerged as the most significant contributor to GHG emissions, followed closely by the Chemical and Manufacturing Industries. This discovery is crucial, as it directs our focus to the sectors where mitigation efforts could be most impactful. We further explored the types of gasses emitted by each industry to understand the distribution of different GHGs across sectors. Counting occurrences of each (Industry, Gas_Type) combination in our dataset, we created a bar plot that provided a clear depiction of the dominant gas types in each sector. Notably, we found that CO2 emissions were significantly prominent in the Energy sector, while other industries exhibited varied emission patterns of greenhouse gasses, including CH4, N2O, and F-gasses. This detailed breakdown is key to tailoring specific emission reduction strategies for each industry, considering their unique emission profiles.





6. **Smart questions 2:**

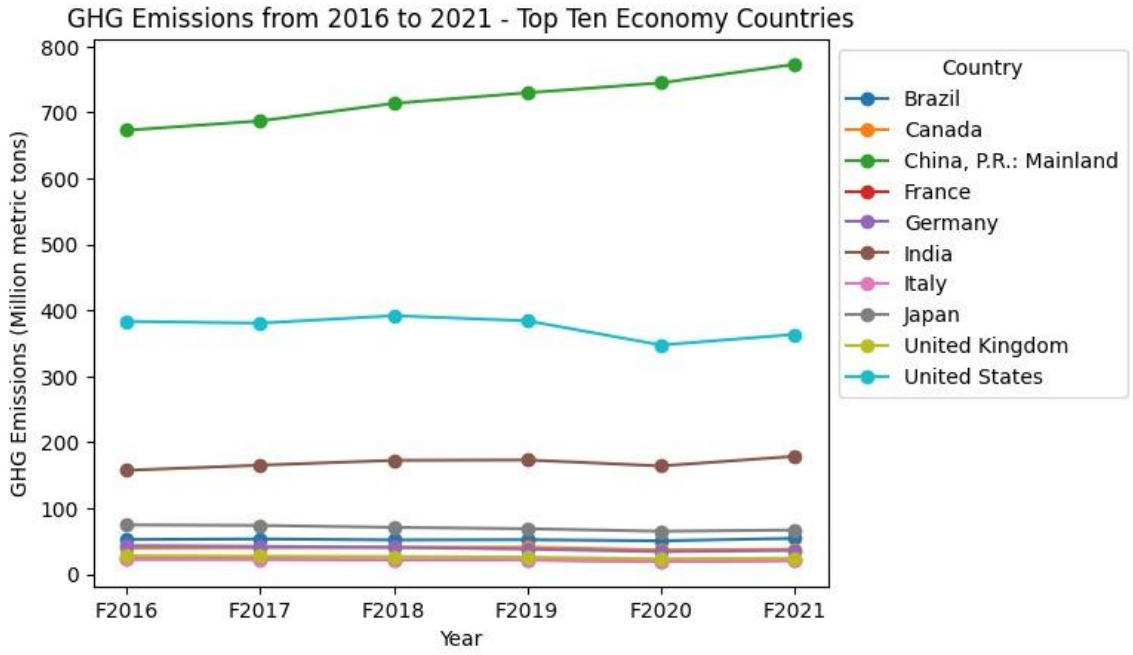
Which industries contribute the most to greenhouse gas (GHG) emissions, in the dataset.

Which gas type is emitted most in every Industry?

Trend Analysis Among Top Economies

To gain insights into the emission trends of the world's top economies, we narrowed our focus to the top ten economic countries, such as the United States, China, Japan, Germany, and India. This analysis involved calculating the mean GHG emissions for each year from 2016 to 2021 for these countries. By plotting these mean values in a line graph, we observed distinct trends in emissions. We noted a consistent increase in emissions from China, highlighting the challenges associated with balancing economic growth and environmental sustainability. In

contrast, the United States displayed a trend of variable emissions, with a notable downward trajectory beginning in 2019, suggesting the implementation of effective measures to curb emissions. European countries like Germany, France, and the United Kingdom exhibited a slight but consistent decline in emissions, reflecting their ongoing commitment to environmental regulations and policies. The technical aspect of this analysis involved meticulously selecting relevant columns for each country and year, grouping the data by country, and calculating the mean emissions. The use of Python's data visualization libraries enabled us to create an informative line graph that clearly illustrated these trends. The observed patterns provide essential insights into the global emissions landscape, emphasizing the need for targeted policy interventions in these key economies to address the challenges posed by GHG emissions.

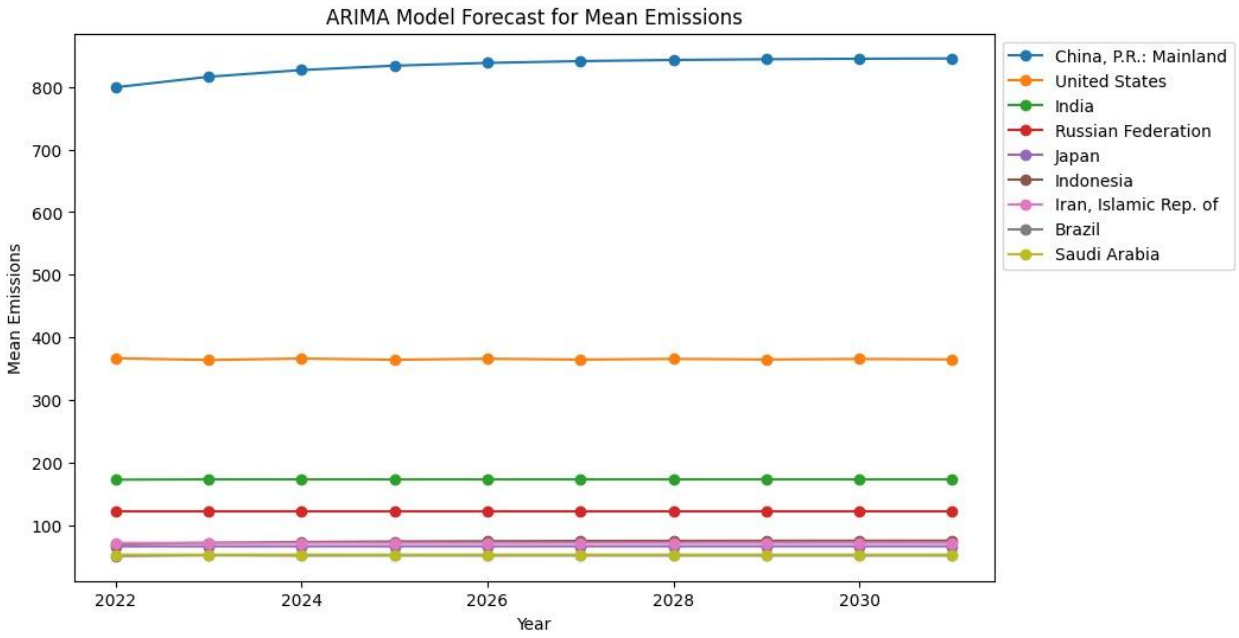


7. Data Modeling

Forecasting future GHG emissions is crucial for understanding the potential impacts of climate change and shaping mitigation strategies. We employed the ARIMA (AutoRegressive Integrated Moving Average) model, renowned for its efficacy in time series forecasting, to project future emissions. Our focus was on the top emitting countries, using mean emissions data from 2016 to 2021 as a foundation.

The ARIMA model fitting involved technical steps such as determining the appropriate order based on Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analysis. Opting for an ARIMA(1,1,1) model, we forecasted emissions for the upcoming decade (2022-2032). The model's forecasts, visualized through line graphs, suggest that countries like China are likely to continue their upward emission trends, while others may experience more stable patterns. This projection is vital for understanding environmental impacts and assisting in policymaking.

The resulting forecasts indicate that under a business-as-usual scenario, most of these countries might maintain or even escalate their emission levels. This trend poses a substantial challenge to global efforts in climate change mitigation, emphasizing the need for urgent and effective policy interventions.



8. **Smart questions 3:**

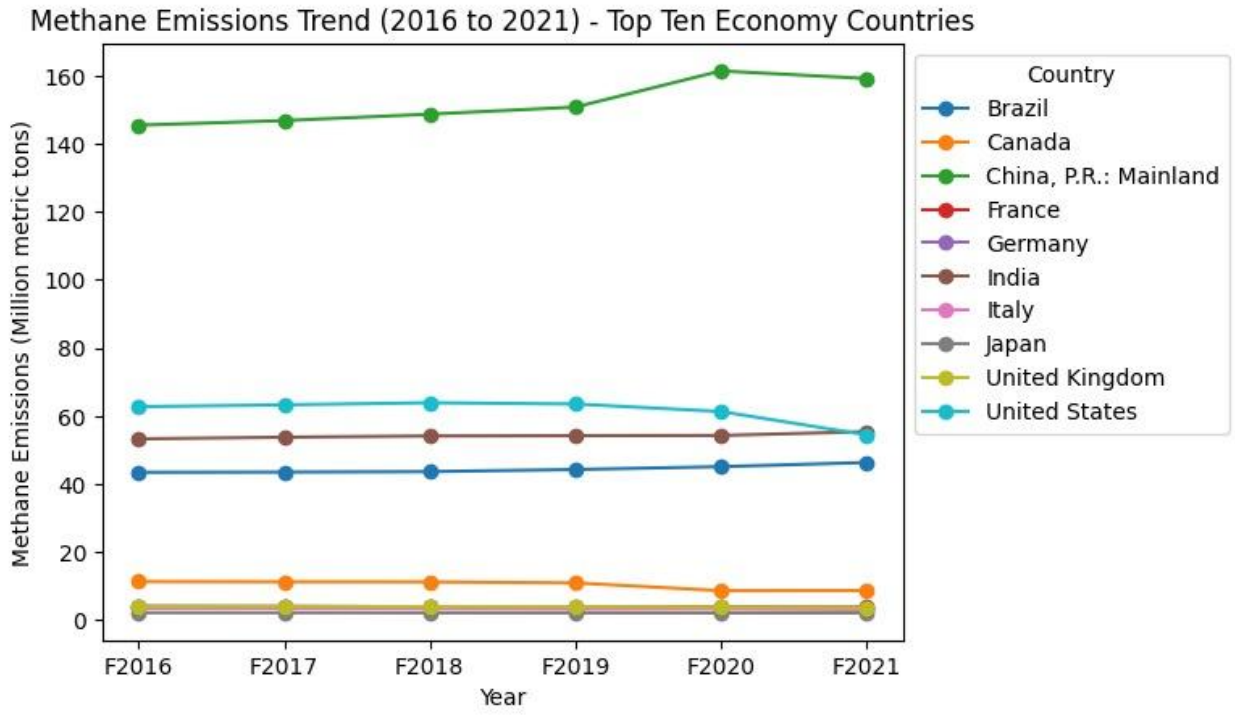
Among all the countries examined in the dataset, which ones have achieved a reduction in methane emissions between 2016 and 2021 indicating efforts in mitigation?

Methane Emissions Mitigation Efforts

Our analysis extended to evaluating methane emission mitigation efforts among the top ten economic countries. Given methane's high global warming potential, understanding its emission trends is crucial. We analyzed the mean emissions for each country from 2016 to 2021, focusing specifically on methane.

The line graph displaying methane emission trends over these years revealed varying patterns across countries. Some nations demonstrated a reduction in methane emissions, signaling successful mitigation strategies. This outcome is particularly encouraging, as it underscores the effectiveness of targeted policies and initiatives in reducing one of the most

potent greenhouse gasses. These trends are a positive indication of the global commitment to mitigating climate change impacts.



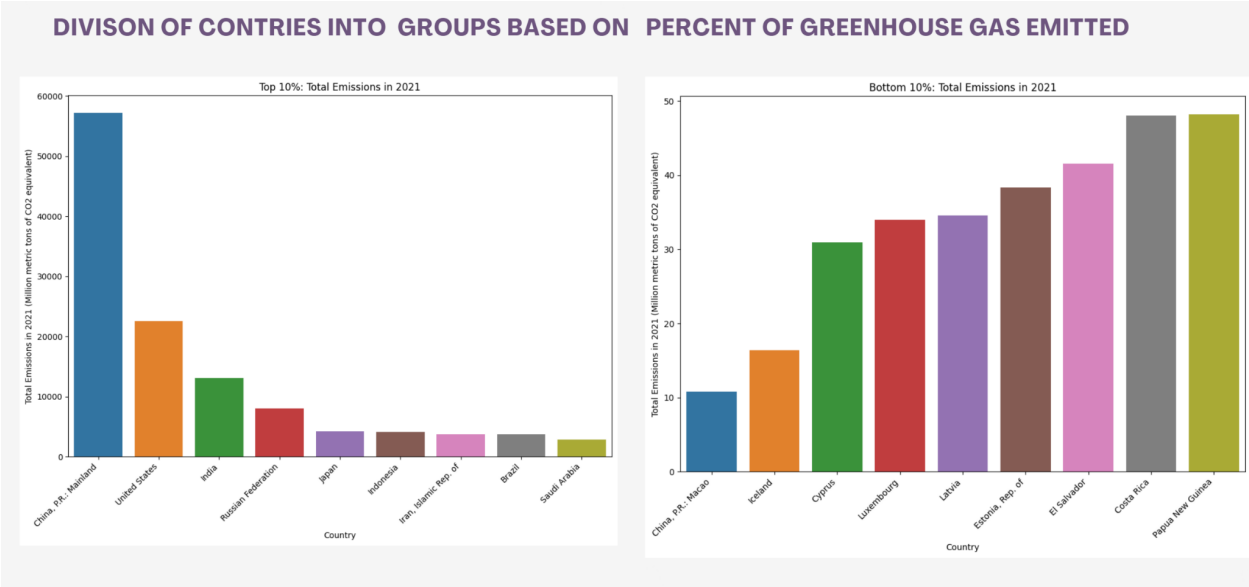
9. **Smart questions 4:**

In terms of emissions, how do the top 10% highest emitting countries, in 2021, compare to the bottom 10% and what factors could explain this discrepancy?

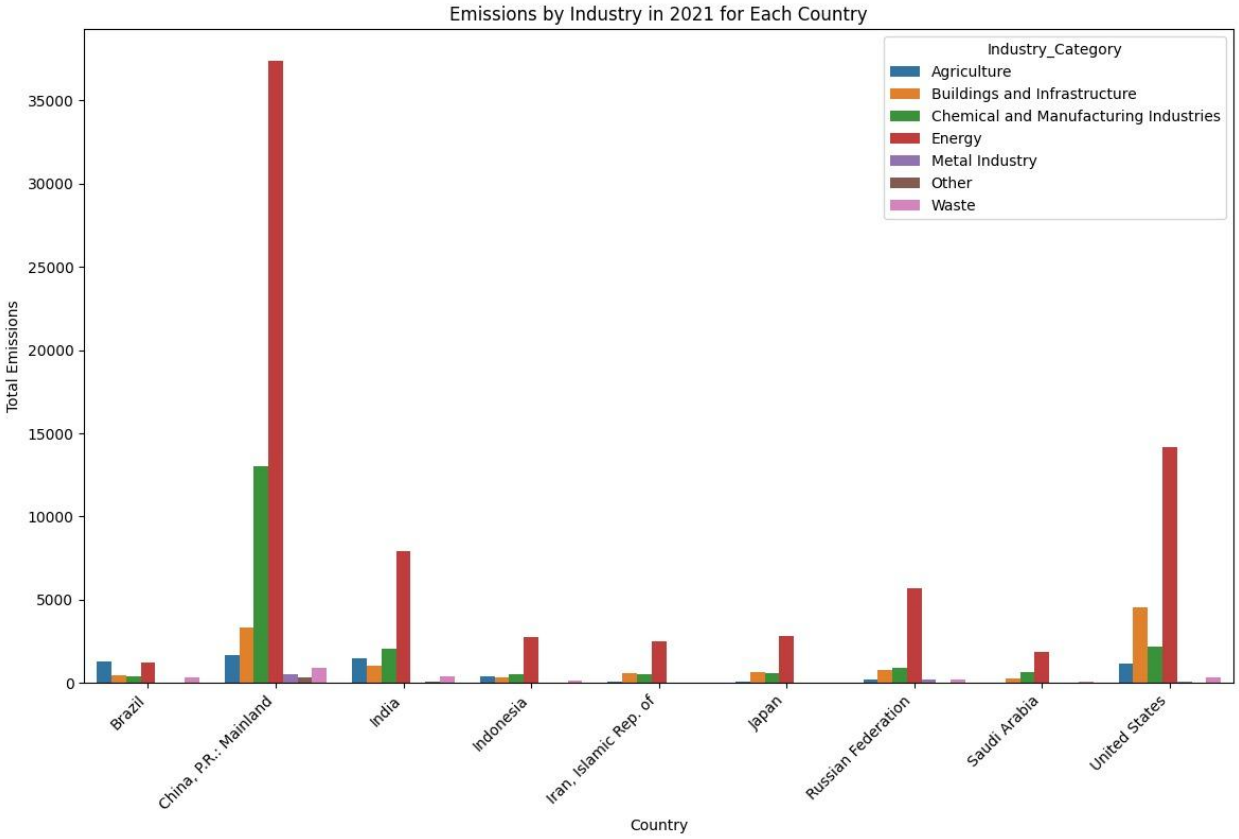
Emissions Disparity Between Countries

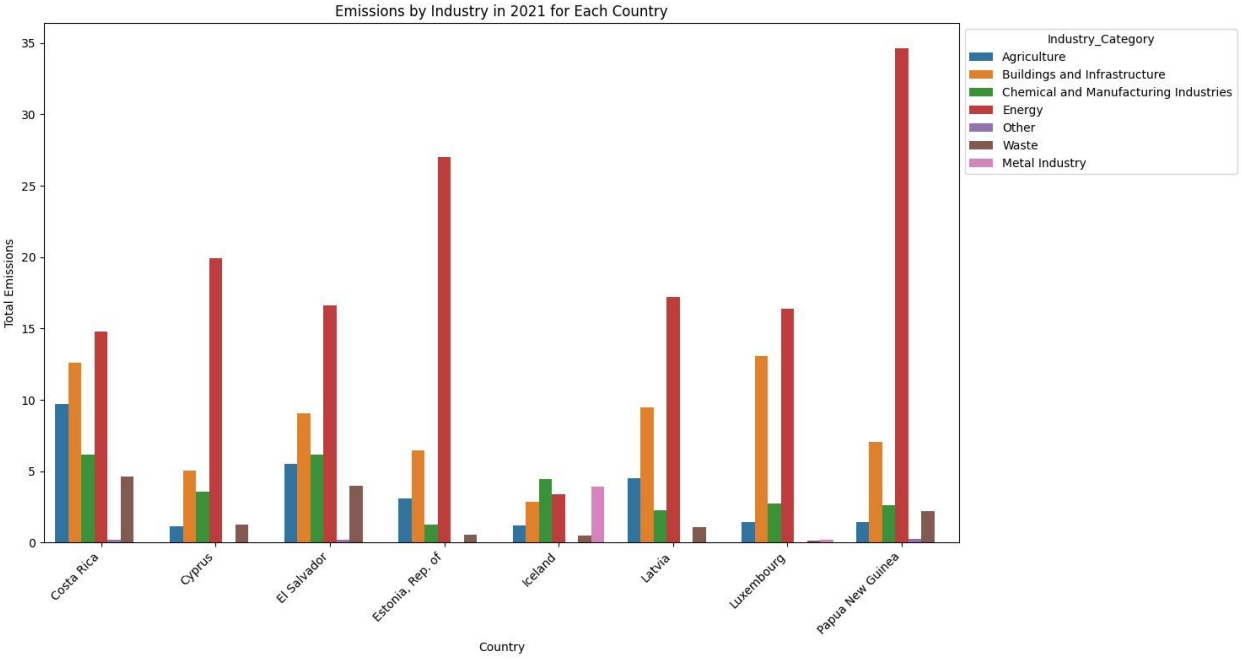
Exploring the disparity in GHG emissions between countries is key to informing global climate policy. We compared the total emissions in 2021 of the top 10% and bottom 10% emitting countries. This analysis was based on computing the total emissions for each country in

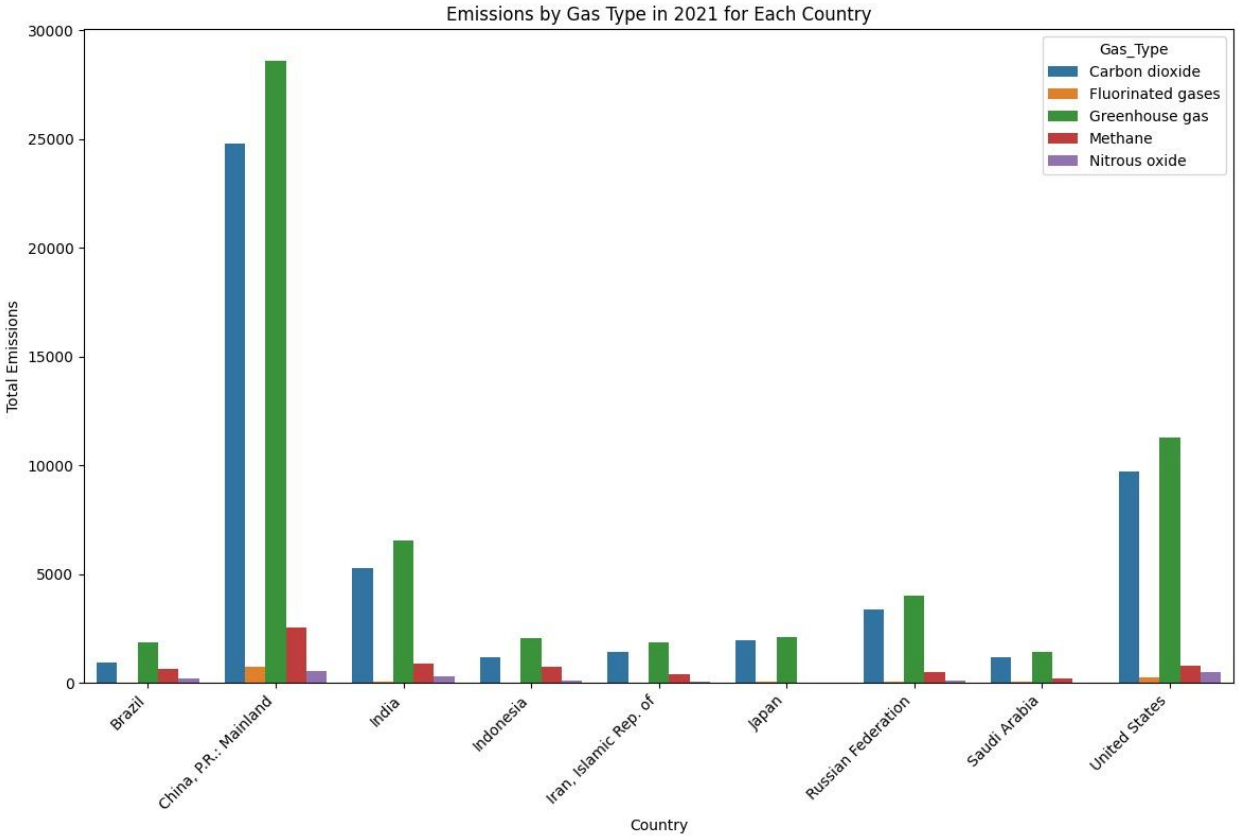
2021 and identifying the highest and lowest emitters.

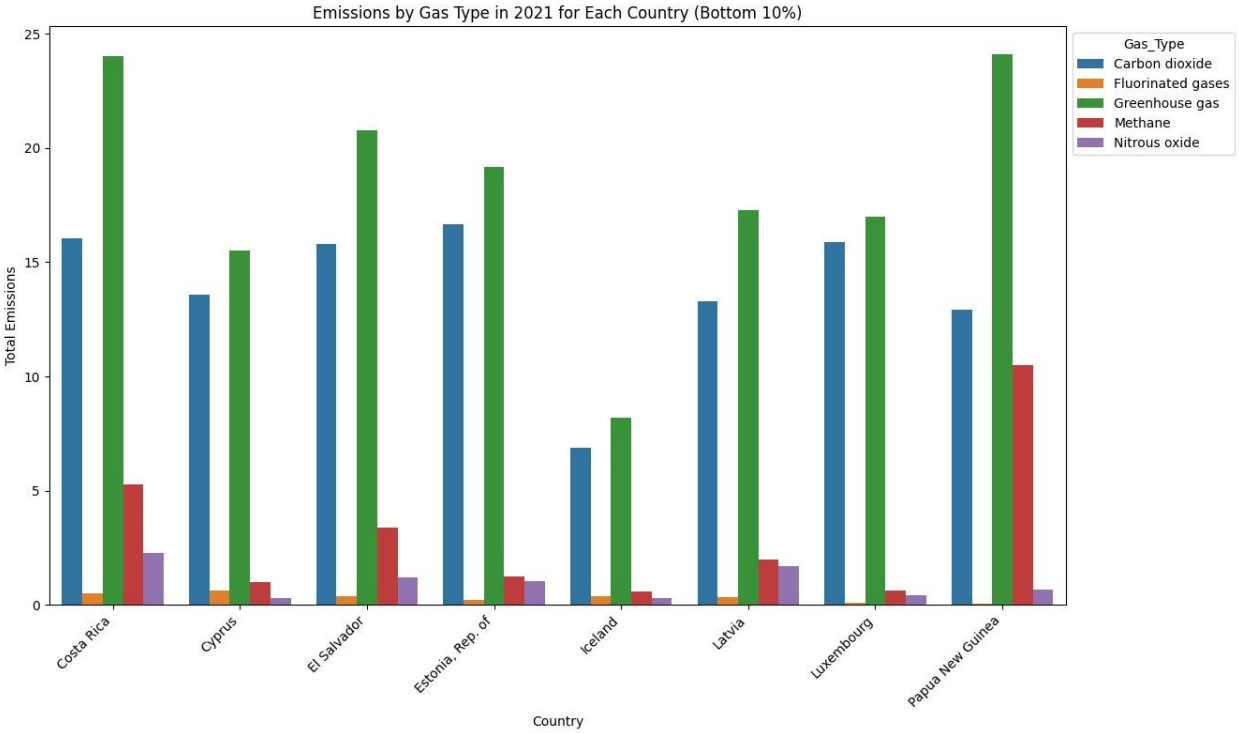


To quantify the differences between these groups, we conducted statistical tests, including independent t-tests and Kolmogorov-Smirnov tests, on the emission data. The t-test focused on the disparity in total emissions, yielding a significant difference between the top and bottom groups, while the Kolmogorov-Smirnov test assessed the distribution of industries and gas types, showing no significant difference in these aspects between the groups. These results not only provide quantitative evidence of the substantial variance in emissions contributions globally but also highlight the consistent distribution of industries and gas types across different emission levels.


Emissions by Industry in 2021 for Each Country
Industry_Category
Agriculture
Buildings and Infrastructure
Chemical and Manufacturing Industries
Energy
Metal Industry
Other
Waste
Total Emissions
0
5000
10000
15000
20000
25000
30000
35000
Brazil
China, P.R.: Mainland
India
Indonesia
Iran, Islamic Rep. of
Japan
Russian Federation
Saudi Arabia
United States
Country


Emissions by Industry in 2021 for Each Country
Industry_Category
Agriculture
Buildings and Infrastructure
Chemical and Manufacturing Industries
Energy
Other
Waste
Metal Industry
Total Emissions
0
5
10
15
20
25
30
35
Costa Rica
Cyprus
El Salvador
Estonia, Rep. of
Iceland
Latvia
Luxembourg
Papua New Guinea
Country


Emissions by Gas Type in 2021 for Each Country
Gas_Type
Carbon dioxide
Fluorinated gases
Greenhouse gas
Methane
Nitrous oxide
Total Emissions
0
5000
10000
15000
20000
25000
30000
Brazil
China, P.R.: Mainland
India
Indonesia
Iran, Islamic Rep. of
Japan
Russian Federation
Saudi Arabia
United States
Country


Emissions by Gas Type in 2021 for Each Country (Bottom 10%)
Gas_Type
Carbon dioxide
Fluorinated gases
Greenhouse gas
Methane
Nitrous oxide
Total Emissions
0
5
10
15
20
25
Costa Rica
Cyprus
El Salvador
Estonia, Rep. of
Iceland
Latvia
Luxembourg
Papua New Guinea
Country

10. CONCLUSION

This comprehensive analysis of global greenhouse gas emissions from 2016 to 2021 has yielded several key insights:

- Sectoral Contributions: The Energy sector emerged as the largest contributor to GHG emissions, underscoring the need for focused mitigation strategies in this area. The sector-specific analysis highlighted the predominant role of CO2 emissions, particularly in energy-intensive industries.
- Emission Trends in Major Economies: The trend analysis revealed a continuous increase in emissions from China and fluctuating trends in the United States. These patterns reflect the varying impacts of economic growth and environmental policies in these countries.
- Future Emissions Projections: Forecasting with the ARIMA model indicated that emissions in top-emitting countries are likely to maintain or increase under a business-as-usual scenario. This finding points to the necessity for more aggressive policies and innovations in green technology to alter these trajectories.
- Methane Mitigation Successes: The analysis of methane emissions showcased successful mitigation efforts in several top economies, especially from 2020 to 2021. This is a positive sign that targeted strategies can effectively reduce emissions of potent greenhouse gasses.
- Disparities Between Countries: The stark differences in emissions between the top 10% and bottom 10% emitting countries highlight the unequal contributions to global GHG emissions. This disparity emphasizes the need for tailored approaches in climate policy, accounting for different countries' diverse economic and industrial profiles.

The findings from our analysis provide a nuanced understanding of global GHG emissions, their sources, and trends. The energy sector emerged as the largest contributor, and specific countries, notably, China, showed significant increases in emissions. The Linear Regression model, with an R-squared value of 0.997, demonstrates the strong predictive power of our approach in forecasting 2021 emissions based on industry categories, gas types, and historical data. The ARIMA model's predictions highlight the need for urgent and effective policy measures, particularly in top-emitting countries, to curb the rising trend of GHG emissions. Our study underscores the importance of targeted mitigation strategies and the role of international cooperation in addressing the global challenge of climate change.

11. REFERENCES

Schneider, S. H. (1989). The greenhouse effect: science and policy. *Science, 243*(4892), 771-781.

Kumar, V., & Garg, M. L. (2018). Predictive analytics: a review of trends and techniques. *International Journal of Computer Applications, 182*(1), 31-37.