UNDERSTANDING GLOBAL GREENHOUSE GAS EMISSIONS THROUGH DATA SCIENCE TECHNIQUES

MOHAMMED RAZA ASFAK CHIDIMAR

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	MOHAMMED RAZA ASFAK CHIDIMAR			
Student's Matric No.	:	MCS231004	Academic Session	:	2023/2024
Date of Birth	:	23/09/1999	UTM Email	:	razaasfak@graduate.utm.my
Choose an item. Title	:	UNDERSTANDING GLOBAL GREENHOUSE GAS EMISSIONS THROUGH DATA SCIENCE TECHNIQUES			

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name MOHAMMED RAZA ASFAK CHIDIMAR
Date : 18/07/2024

Approved by Supervisor(s)

Signature of Supervisor I:

Full Name of Supervisor I
PROF. MADYA DR MOHD SHAHIZAN OTHMAN

Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:

Librarian

Jabatan Perpustakaan UTM,

Universiti Teknologi Malaysia,

Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: UNDERSTANDING GLOBAL GREENHOUSE GAS EMISSIONS THROUGH DATA SCIENCE TECHNIQUES

AUTHOR'S FULL NAME:MOHAMMED RAZA ASFAK CHIDIMAR

Please be informed that the above-mentioned thesis titled _____ should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE:

NAME:

ADDRESS OF SUPERVISOR: PROF. MADYA DR MOHD SHAHIZAN OTHMAN

"Choose an item. hereby declare that Choose an item. have read this Choose an item. and in Choose an item.
opinion this Choose an item. is sufficient in term of scope and quality for the award of the degree of Choose an item."

Signature	:	_____
Name of Supervisor I	:	PROF. MADYA DR. MOHD. SHAHIZAN OTHMAN
Date	:	18 JUNE 2024

Signature	:	_____
Name of Supervisor II	:	
Date	:	

Signature	:	_____
Name of Supervisor III	:	
Date	:	

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

UNDERSTANDING GLOBAL GREENHOUSE GAS EMISSIONS THROUGH DATA SCIENCE TECHNIQUES

MOHAMMED RAZA ASFAK CHIDIMAR

A project report submitted in partial fulfilment of the requirements for the award of the degree of Master of Science (Data Science)

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2024

DECLARATION

I declare that this project report entitled *"Understanding Global Greenhouse Gas Emissions Through Data Science Techniques"* is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	MOHAMMED RAZA ASFAK CHIDIMAR
Date	:	18 JUNE 2024

ACKNOWLEDGEMENT

In preparing this thesis, I was fortunate to interact with many individuals who significantly contributed to my understanding and insights. I would like to express my sincere gratitude to my main thesis supervisor, Professor Madya Dr. Mohd Shahizan Othman, for his encouragement, guidance, critiques, and friendship. His support was instrumental in the completion of this work.

I am also grateful to my fellow postgraduate students for their unwavering support and to all my colleagues and others who have provided assistance at various stages. Although I cannot name everyone in this limited space, your contributions are deeply appreciated.

Finally, I am immensely thankful to my family for their constant support and encouragement. Your belief in me has been my driving force.

ABSTRACT

This research addresses the critical issue of rising global greenhouse gas (GHG) emissions, a major contributor to climate change and environmental degradation. Despite international efforts, emissions have continued to increase, driven by factors such as industrial activities, transportation, energy production, and agricultural practices. This study employs advanced data science techniques, including exploratory data analysis (EDA), predictive modelling, and spatial analysis, to provide a comprehensive examination of GHG emissions from 1970 to 2019. Utilizing data from the Emissions Database for Global Atmospheric Research (EDGAR) and other reputable sources, the research aims to identify temporal trends, sectoral contributions, and regional differences in emissions. The objectives of this project include analysing seasonal and long-term patterns in GHG emissions, evaluating the impact of different sectors, and developing data-driven mitigation strategies. By leveraging detailed emissions data and sophisticated analytical methods, the study seeks to inform policy development and enhance climate change mitigation efforts. The findings are expected to provide valuable insights for policymakers, stakeholders, and international organizations, supporting informed decision-making and promoting sustainable development. This research contributes to the field of environmental science by offering robust predictive models and actionable recommendations to address the escalating challenge of GHG emissions, ultimately aiming to foster a more sustainable and resilient future.

ABSTRAK

Kajian ini menangani isu kritikal peningkatan gas rumah hijau (GHG) global yang merupakan penyumbang utama kepada perubahan iklim dan degradasi alam sekitar. Walaupun terdapat usaha antarabangsa, pelepasan GHG terus meningkat, didorong oleh aktiviti perindustrian, pengangkutan, pengeluaran tenaga, dan amalan pertanian. Kajian ini menggunakan teknik sains data yang maju, termasuk analisis data eksploratori (EDA), pemodelan ramalan, dan analisis spatial, untuk memberikan pemeriksaan menyeluruh terhadap pelepasan GHG dari tahun 1970 hingga 2019. Dengan menggunakan data dari Emissions Database for Global Atmospheric Research (EDGAR) dan sumber lain yang bereputasi, kajian ini bertujuan untuk mengenal pasti trend temporal, sumbangan sektoral, dan perbezaan serantau dalam pelepasan GHG. Objektif projek ini termasuk menganalisis corak bermusim dan jangka panjang dalam pelepasan GHG, menilai impak pelbagai sektor, dan membangunkan strategi mitigasi berasaskan data. Dengan memanfaatkan data pelepasan yang terperinci dan kaedah analisis yang sofistikated, kajian ini bertujuan untuk memaklumkan pembangunan dasar dan meningkatkan usaha mitigasi perubahan iklim. Penemuan kajian ini diharapkan dapat memberikan pandangan berharga kepada pembuat dasar, pemegang kepentingan, dan organisasi antarabangsa, menyokong pembuatan keputusan yang berinformasi dan mempromosikan pembangunan lestari. Kajian ini menyumbang kepada bidang sains alam sekitar dengan menawarkan model ramalan yang kukuh dan cadangan tindakan untuk menangani cabaran peningkatan pelepasan GHG, dengan tujuan akhirnya untuk membina masa depan yang lebih lestari dan berdaya tahan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARIMA	-	AutoRegressive Integrated Moving Average
CO2	-	Carbon Dioxide
CNNs	-	Convolutional Neural Networks
EDA	-	Exploratory Data Analysis
GHG	-	Greenhouse Gas
GIS	-	Geographic Information System
IEA	-	International Energy Agency
LSTM	-	Long Short-Term Memory
MAE	-	Mean Absolute Error
MLP	-	Multilayer Perceptrons
OMP	-	Orthogonal Matching Pursuit Regression
RNN	-	Recurrent Neural Networks
RMSE	-	Root Mean Square Error
SVM	-	Support Vector Machines
TFT	-	Temporal Fusion Transformers

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The escalation of global greenhouse gas (GHG) emissions is a major contributor to climate change, impacting environmental health, economic stability, and societal well-being. Much like the fluctuations in grocery prices affect the Malaysian economy, rising GHG emissions shape global climate policies and national frameworks. Key sources of these emissions include industrial activities, transportation, energy production, and agricultural practices, all of which drive global warming and environmental degradation. Despite numerous international efforts to reduce emissions, such as the Paris Agreement, GHG levels continue to rise, presenting significant challenges to achieving climate goals.

Recent data highlights that global GHG emissions have been increasing consistently over the past few decades, with notable contributions from both developed and developing nations (Minx et al., 2021; Brookings, 2023). This ongoing growth is influenced by a complex interplay of factors, including economic expansion, population growth, and technological advancements. The International Energy Agency (IEA) reported that CO2 emissions reached an unprecedented high in 2023, emphasizing the urgent need for effective mitigation strategies (IEA, 2023).

Understanding the sources and trends of GHG emissions is essential for developing effective mitigation strategies. Advanced data science techniques, such as exploratory data analysis (EDA), predictive modeling, and spatial analysis, offer valuable insights into emissions data. For example, Mahajan et al. (2022) used EDA and time-series forecasting to identify the causes of global warming and predict CH4 concentrations with high accuracy. Similarly, Wang (2022) demonstrated the potential

of machine learning models in forecasting global GHG emissions per capita, showcasing the significant role of data science in environmental research.

This project aims to utilize these advanced data science techniques to conduct a comprehensive analysis of GHG emissions from 1970 to 2019. By employing data from the Emissions Database for Global Atmospheric Research (EDGAR) and other reputable sources, the study will identify temporal trends, sectoral contributions, and regional differences in emissions. The objectives include analyzing seasonal and long-term patterns in GHG emissions, evaluating the impact of various sectors, and proposing data-driven mitigation strategies.

Through this approach, the project seeks to inform policy development and enhance climate change mitigation efforts, providing valuable insights for policymakers, stakeholders, and international organizations. The findings are expected to contribute to the broader discourse on sustainable development and environmental resilience, ultimately aiming to foster a more sustainable and resilient future.

1.2 Problem Background

Global greenhouse gas (GHG) emissions present a persistent and critical challenge, impacting environmental health, economic stability, and societal well-being. The continuous rise in GHG emissions is a major driver of climate change, resulting in severe consequences such as global warming, sea-level rise, and increased frequency of extreme weather events. This escalation has drawn concern from policymakers, researchers, and the public, underscoring the urgent need for comprehensive analysis and effective mitigation strategies.

Despite international agreements like the Paris Agreement aimed at curbing emissions, global GHG levels have continued to rise. Industrial activities, energy production, transportation, and agricultural practices collectively drive the majority of

emissions (Minx et al., 2021). The complexity of these sources necessitates a detailed and nuanced approach to effectively address the underlying causes.

Recent data emphasizes the urgency of this issue. The International Energy Agency (IEA) reported that CO2 emissions reached a record high in 2023, driven by economic growth, population increase, and technological advancements (IEA, 2023). Additionally, Brookings (2023) highlighted that both developed and developing countries contribute to this upward trend, complicating efforts to achieve global climate targets. Understanding the sources and trends of GHG emissions is essential for developing effective mitigation strategies. Traditional methods of emissions analysis often fall short in capturing the complexity needed for accurate predictions and policy-making. Therefore, advanced data science techniques, such as exploratory data analysis (EDA), predictive modeling, and spatial analysis, offer transformative approaches to this challenge.

Environmental studies have demonstrated the efficacy of these techniques. For instance, Mahajan et al. (2022) used EDA and time-series forecasting to understand the causes of global warming and predict CH4 concentrations with high accuracy. Similarly, Wang (2022) showcased the potential of machine learning models in forecasting global GHG emissions per capita, highlighting the importance of integrating diverse data sources and sophisticated analytical methods.

In response to these challenges, this project aims to leverage advanced data science techniques to provide a comprehensive analysis of GHG emissions from 1970 to 2019. By utilizing data from the Emissions Database for Global Atmospheric Research (EDGAR) and other reputable sources, the study seeks to identify temporal trends, sectoral contributions, and regional differences in emissions. This approach will enable the development of data-driven mitigation strategies, informing policy development and enhancing climate change mitigation efforts.

Through this detailed analysis, the project aims to contribute valuable insights to the global discourse on sustainable development and environmental resilience. By addressing the limitations of traditional methods and incorporating advanced data

science techniques, the research will offer stakeholders more accurate and actionable information for navigating and responding to the complex dynamics of global GHG emissions.

1.3 Problem Statement

This study aims to tackle the critical issue of accurately forecasting global greenhouse gas (GHG) emissions, which is essential for policymakers, researchers, and environmental organizations to make informed decisions and implement effective climate strategies. Despite extensive international efforts to curb emissions, global GHG levels continue to rise, driven by industrial activities, transportation, energy production, and agricultural practices (Minx et al., 2021). The complexity of these sources and the lack of a comprehensive predictive model that incorporates various influential factors hinder accurate projections of future emissions. Traditional methods of emissions forecasting often fall short due to their inability to integrate crucial external elements like economic indicators, technological advancements, and regional differences (IEA, 2023). This limitation restricts the ability of stakeholders to proactively manage and mitigate the impact of rising emissions. Therefore, the goal of this research is to develop a robust predictive model based on advanced data science techniques, including exploratory data analysis (EDA), predictive modeling, and spatial analysis, utilizing data from the Emissions Database for Global Atmospheric Research (EDGAR) and other reputable sources. By creating this comprehensive predictive model, the study seeks to provide valuable insights into the temporal trends, sectoral contributions, and regional differences in GHG emissions. This model will enable stakeholders to better understand the dynamics of emissions and implement more effective mitigation strategies. Ultimately, the research aims to contribute to global efforts in combating climate change by offering a reliable tool for forecasting GHG emissions, supporting informed decision-making, and promoting sustainable development.

1.4 Objectives

The proposed project aims to achieve the following objectives:

I. To To analyse global greenhouse gas (GHG) emissions using advanced data science techniques, focusing on key sectors such as industry, transportation, energy production, and agriculture.

II. To develop predictive models for GHG emissions using machine learning algorithms, including random forests, neural networks, and hybrid models, to forecast future emissions trends.

III. Develop data-driven strategies for mitigating GHG emissions based on comprehensive analysis and predictive modelling.

1.5 Gap Analysis

There is a substantial body of literature examining the application of data science techniques to forecast various environmental phenomena, including greenhouse gas (GHG) emissions. Studies like those by Erfanian et al. (2022) have utilized machine learning to predict air quality indices based on environmental and meteorological data, demonstrating the potential of these techniques in environmental forecasting. Similarly, Agbo et al. (2022) reviewed the use of machine learning methods for predicting energy consumption and emissions, highlighting their relevance in energy and environmental management. These studies underscore the effectiveness of machine learning in various domains, including environmental science.

In the specific context of GHG emissions, Chen et al. (2021) conducted comparative experiments on multiple machine learning algorithms to predict CO2 emissions from different industrial sectors, showcasing the utility of these techniques in environmental monitoring. Additionally, Ho et al. (2020) applied predictive

modeling, including support vector machines and random forests, to forecast emissions from transportation sources, indicating the applicability of these methods in transportation emissions prediction.

Overall, the existing literature highlights the potential of data science techniques in predicting GHG emissions across various sectors. However, there is a notable research gap in the application of these advanced techniques specifically for comprehensive GHG emissions forecasting that incorporates a wide range of influential factors, including economic indicators, technological advancements, and regional disparities. This project aims to bridge this gap by developing tailored predictive models that leverage data science techniques to provide a holistic and precise forecast of GHG emissions. The study will utilize historical emissions data, along with various external factors, to refine the accuracy of predictions and support effective climate change mitigation strategies. By addressing this gap, the research seeks to enhance the predictive capabilities in the field of GHG emissions and contribute to more informed and proactive environmental policy-making.

1.6 Scope

The scope of this study, "Understanding Global Greenhouse Gas Emissions Through Data Science Techniques," is extensive and multifaceted. It involves a thorough analysis of historical GHG emissions data from 1970 to 2019, sourced from the Emissions Database for Global Atmospheric Research (EDGAR) and other reputable sources. The project encompasses meticulous data collection, preprocessing, and exploratory analysis to identify data patterns, outliers, and correlations. It includes feature engineering by integrating external factors such as economic indicators, technological advancements, and regional disparities into predictive models to enhance their accuracy and robustness. The development, training, and fine-tuning of advanced data science techniques, including machine learning algorithms and statistical models, are central to this study. Validation and evaluation of these models ensure their reliability. Insights derived from these predictive models will be translated into actionable information for stakeholders, assisting policymakers, researchers, and

environmental organizations in making informed decisions. The project's scope also addresses limitations, provides recommendations for future enhancements, and acknowledges the dynamic nature of factors influencing GHG emissions within the constraints of available data and model complexity.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview of Greenhouse Gas Emissions

Greenhouse gas (GHG) emissions are a critical factor contributing to global warming and climate change. The analysis of these emissions is essential for understanding their impact on the environment and for developing strategies to mitigate their effects. Data science techniques play a vital role in analyzing GHG emissions data, enabling researchers to identify trends, forecast future emissions, and develop data-driven policies for emission reduction. This section provides an overview of the historical context and trends of GHG emissions and emphasizes the importance of accurate data and advanced analytical methods in addressing climate change.

2.1.1 Historical Context and Trends

The historical analysis of GHG emissions reveals persistent growth over the past decades, with significant contributions from various sectors such as industry, transportation, and agriculture. Mahajan and Jain (2022) conducted an extensive analysis of GHG concentration records, developing forecasting models that achieved high accuracy, particularly for methane (CH4) concentrations. Their study highlights the critical role of GHGs in global warming and the necessity for accurate predictive models to inform policy decisions.

Minx et al. (2021) compiled a comprehensive dataset of global GHG emissions from 1970 to 2018, providing valuable insights into the long-term trends of emissions. Their research shows that despite numerous international agreements and efforts to reduce emissions, the overall global emissions have continued to rise. This persistent increase underscores the challenges in achieving significant reductions and the importance of continued monitoring and analysis.

The compilation of data by Our World in Data (2023) further emphasizes the differences in the warming impacts of various gases, such as methane and carbon dioxide (CO2). The introduction of the Global Warming Potential (GWP*) metric provides a more accurate representation of the short-lived gases' impact on global warming, which is crucial for developing targeted mitigation strategies (Our World in Data, 2023).

Ciais et al. (2023) presented the GRACED dataset, which offers near-real-time global gridded daily CO2 emissions data with fine spatial resolution. This dataset is instrumental in monitoring emissions trends and understanding the immediate impact of policy changes and economic activities on GHG emissions. The high resolution and timely data provided by GRACED enable more precise and dynamic analysis, which is essential for effective climate action planning. In summary, the historical context and trends of GHG emissions illustrate the persistent challenges in managing global emissions and highlight the need for advanced data science techniques to analyze and predict these trends. The integration of comprehensive datasets and accurate forecasting models is crucial for developing effective policies and strategies to mitigate the impact of GHG emissions on global warming.

2.2 Data Sources and Datasets

Understanding the sources and quality of data is fundamental to conducting a comprehensive analysis of greenhouse gas (GHG) emissions. Reliable datasets enable researchers to accurately track historical trends, analyze current patterns, and develop predictive models that inform policy decisions and mitigation strategies. This section highlights the key datasets and data sources that will be utilized in this study, emphasizing their relevance and importance in the context of GHG emissions research.

2.2.1 Our World in Data (2023)

Our World in Data provides a comprehensive overview of global greenhouse gas emissions, highlighting the differences in warming impacts among various gases,

such as methane (CH4) and carbon dioxide (CO2). The introduction of the Global Warming Potential (GWP*) metric is particularly significant. The GWP* metric offers a more accurate representation of the short-lived gases' impact on global warming, which is essential for developing effective mitigation strategies. The data from this source is crucial for understanding the historical and current trends in GHG emissions and their impacts on global climate patterns. By utilizing this dataset, researchers can better understand the relative contributions of different gases to global warming and devise more targeted reduction strategies (Our World in Data, 2023).

2.2.2 Ciais et al. (2023)

Ciais et al. (2023) presented the GRACED dataset, a near-real-time global gridded daily CO2 emissions dataset. This dataset provides fine spatial resolution, making it invaluable for monitoring emissions trends and understanding the immediate effects of policy changes and economic activities on GHG emissions. The high resolution and timeliness of the GRACED data enable more precise and dynamic analysis, which is crucial for effective climate action planning. The dataset includes CO2 emissions from various sectors, including power, industry, residential consumption, and transportation, offering a detailed view of emissions across different sources. This granularity is essential for identifying key emission sources and evaluating the impact of specific mitigation measures (Ciais et al., 2023).

2.2.3 Comprehensive Data Sources

Minx et al. (2021) compiled a synthetic dataset that covers global, regional, and national greenhouse gas emissions by sector from 1970 to 2018, with an extension to 2019. This dataset includes CO2, CH4, N2O, and fluorinated gases (F-gases), providing detailed insights into the contributions of different sectors to overall emissions. The persistent growth in emissions across all sectors underscores the ongoing challenges in reducing global GHG emissions and highlights the need for robust data to inform policy decisions. The comprehensive nature of this dataset allows for a nuanced analysis of emissions trends and the effectiveness of international agreements aimed at reducing emissions (Minx et al., 2021).

2.2.4 Climate Data from IMF

The dataset available from the International Monetary Fund (IMF) at Climate Data is a crucial resource for this study. This dataset provides comprehensive data on greenhouse gas emissions, climate policies, and economic indicators, allowing for an integrated analysis of the factors influencing emissions. Utilizing this dataset will enable the examination of the relationships between economic activities and GHG emissions, providing insights into how policy changes and economic growth impact emissions levels. The detailed and up-to-date data from the IMF is essential for developing accurate predictive models and formulating effective mitigation strategies. By incorporating economic indicators, this dataset allows for a more comprehensive understanding of the drivers of GHG emissions and the potential impacts of different policy measures.

2.2.5 Importance of Accurate Data

Accurate and comprehensive datasets are essential for understanding and addressing the complexities of global greenhouse gas emissions. The integration of data from multiple sources, including historical records, near-real-time data, and sector-specific emissions, enables a holistic analysis of emissions trends. Advances in remote sensing technologies and satellite observations have further enhanced the precision and coverage of emission data. This comprehensive approach is necessary to develop effective mitigation strategies and inform policy decisions that can significantly impact global efforts to combat climate change. High-quality data ensures that policymakers and researchers can make informed decisions based on reliable evidence, ultimately leading to more effective climate action. Additionally, the use of big data analytics and machine learning algorithms can uncover hidden patterns and predict future emissions scenarios, supporting proactive and adaptive policy measures.

2.3 Predictive Modeling Techniques

Predictive modeling is a cornerstone of data science, enabling researchers to forecast future trends based on historical data. In the context of greenhouse gas (GHG) emissions, predictive models help in understanding future emissions scenarios and in formulating effective mitigation strategies. This section delves into the various predictive modeling techniques used in GHG emissions research, highlighting the strengths and applications of each method.

2.3.1 Statistical Models

Statistical models have long been used to analyze and predict time-series data. Traditional models such as the autoregressive (AR), moving average (MA), and autoregressive integrated moving average (ARIMA) are fundamental in time-series forecasting. These models are particularly useful for their simplicity and effectiveness in linear data analysis.

2.3.1.1 ARIMA Model by Mahajan and Jain (2022)

The Autoregressive Integrated Moving Average (ARIMA) model combines three key components: autoregression (AR), differencing to achieve stationarity (I for integrated), and moving average (MA). Mahajan and Jain (2022) utilized ARIMA models to forecast GHG concentrations, achieving high accuracy in their predictions. The ARIMA model's ability to handle seasonality and trends in the data makes it a suitable choice for forecasting emissions, which often exhibit seasonal patterns and long-term trends (Mahajan and Jain, 2022).

The ARIMA model involves identifying the model parameters (p, d, q) where p is the number of lag observations (autoregressive part), d is the number of times differencing is applied (integrated part), and q is the size of the moving average window (moving average part). The model is then fitted to historical data to estimate these parameters. Once fitted, the model can generate forecasts by predicting future emissions based on the historical data and the estimated parameters. The ARIMA model formula is:

$$y_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \quad (2.1)$$

where y_t is the value at time t, c is a constant, ϵ_t is the white noise error term, ϕ_i are the autoregressive coefficients, and θ_j are the moving average coefficients.

2.3.2 Machine Learning Techniques

With the advent of machine learning, more sophisticated predictive models have been developed. Machine learning algorithms like support vector machines (SVM), random forests (RF), and neural networks have shown superior performance in handling complex, nonlinear data.

2.3.2.1 Random Forests by Wang (2022)

Random Forests (RF) are an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Wang (2022) demonstrated the effectiveness of Random Forests in predicting short-term GHG emissions with high accuracy. This technique is particularly advantageous for its ability to handle large datasets and capture non-linear relationships within the data (Wang, 2022).

The Random Forest algorithm starts with bootstrap sampling, where samples are randomly selected from the training dataset with replacement. For each bootstrap sample, a decision tree is constructed. At each node in the tree, the best split is selected from a random subset of features. This process helps in reducing the variance by averaging out the predictions of multiple trees, thereby increasing the model's overall robustness and accuracy. Finally, the results are aggregated: for regression, the predictions of all trees are averaged, and for classification, majority voting is used. The Random Forest prediction formula is:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x) \tag{2.2}$$

where $\hat{f}(x)$ is the predicted value, $\hat{f}(x)$ is the number of trees, and $f_i(x)$ is the prediction of the i -th tree.

2.3.3 Deep Learning and Hybrid Models

Deep learning models, such as convolutional neural networks (CNNs) and transformer-based models, have revolutionized predictive modeling by offering exceptional capabilities in pattern recognition and feature extraction.

2.3.3.1 Temporal Fusion Transformers (TFT) by Frontiers (2023)

Temporal Fusion Transformers (TFT) are a state-of-the-art deep learning technique for handling time-series data, particularly effective in capturing long-term dependencies and complex patterns. Frontiers (2023) explored the use of TFT for high-resolution GHG emissions forecasting, applying the model to maritime activities. TFT leverages attention mechanisms to focus on the most relevant parts of the data, significantly enhancing prediction accuracy. This model is especially useful for analysing large, multi-dimensional datasets and providing detailed insights into future emission trends (Frontiers, 2023).

The Temporal Fusion Transformer algorithm starts with input data preprocessing to prepare the necessary features. This data is then processed through Long Short-Term Memory (LSTM) units to handle sequential data. Attention mechanisms are applied to capture relevant features across time steps, and predictions are generated using fully connected layers.

2.3.3.2 Hybrid Models by Nature (2023)

Hybrid models that combine traditional statistical methods with machine learning algorithms offer a balanced approach to predictive modelling. Nature (2023) discussed the integration of ARIMA with machine learning techniques to improve the

robustness and accuracy of GHG emissions forecasts. These hybrid models capitalize on the strengths of both approaches, making them versatile and powerful tools for environmental data analysis (Nature, 2023).

The hybrid model algorithm starts with fitting an ARIMA model to capture linear patterns and extracting the residuals. A machine learning model (e.g., Random Forests) is then trained on these residuals to capture non-linear patterns. The final predictions are obtained by combining the predictions from both the ARIMA and machine learning components. The hybrid model prediction formula is:

$$\widehat{y_t} = \mathrm{ARIMA}(x_t) + \mathrm{ML}(x_t) \tag{2.3}$$

where y_t is the final prediction, $\mathrm{ARIMA}(x_t)$ is the ARIMA model prediction, and $\mathrm{ML}(x_t)$ is the machine learning model prediction.

2.3.4 Neural Networks by Julian et al. (2023)

Neural networks, including multilayer perceptrons (MLP) and recurrent neural networks (RNN), have been widely adopted for time-series prediction tasks. Julian et al. (2023) used neural networks to predict CO2 emissions from new cars, demonstrating the potential of these models in understanding the impact of private transportation on emissions. The study highlights the capabilities of neural networks in capturing complex relationships in the data and providing accurate forecasts (Julian et al., 2023).

The neural network algorithm starts with defining the input features and constructing multiple hidden layers with activation functions (e.g., ReLU). The output layer is defined with an appropriate activation function (e.g., linear for regression). The model is trained using backpropagation to minimize the loss function. The neural network prediction formula is:

$$y = \sigma(W \cdot x + b) \tag{2.4}$$

where y is the output, σ is the activation function, W is the weight matrix, x is the input vector, and b is the bias vector.

2.3.5 Orthogonal Matching Pursuit Regression (OMP) by Zhe (2022)

Orthogonal Matching Pursuit Regression (OMP) is a machine learning algorithm known for its high accuracy in handling complex datasets. Zhe (2022) demonstrated the effectiveness of OMP in forecasting global GHG emissions per capita. This model is particularly useful for its ability to handle large, multidimensional datasets and make precise predictions, showcasing its potential in environmental data analysis (Zhe, 2022).

The Orthogonal Matching Pursuit Regression algorithm starts with initializing the residuals with the original target values. Features are iteratively selected that most reduce the residuals. The residuals are recalculated by removing the contribution of the selected feature. This process continues until a stopping criterion is met. The Orthogonal Matching Pursuit Regression formula is:

$$\beta = \arg\min_{\beta} | y - X\beta|_2 \tag{2.5}$$

where y is the target vector, X is the feature matrix, and β is the coefficient vector.

2.3.6 Importance of Predictive Modelling

Predictive modeling is crucial for anticipating future emissions and understanding the potential impacts of various mitigation strategies. Accurate predictions enable policymakers and researchers to make informed decisions, allocate resources efficiently, and implement timely interventions. The continuous development and refinement of predictive models are essential for advancing our understanding of GHG emissions and for driving effective climate action.

Table 2.1 Different Models Used in GHG Emissions Prediction

Study	Model(s) Used	Key Findings
Mahajan and Jain (2022)	ARIMA	Achieved high accuracy in forecasting GHG concentrations, highlighting the model's ability to handle seasonal and trend patterns in emissions data.
Wang (2022)	Random Forests (RF)	Demonstrated high accuracy in short-term GHG emissions prediction, effectively handling large datasets and capturing non-linear relationships.
Frontiers (2023)	Temporal Fusion Transformers (TFT)	Applied TFT to maritime GHG emissions forecasting, leveraging attention mechanisms to enhance prediction accuracy in large, multi-dimensional datasets.
Nature (2023)	Hybrid models combining ARIMA and machine learning	Improved robustness and accuracy of GHG emissions forecasts by integrating traditional statistical methods with advanced machine learning techniques.
Julian et al. (2023)	Neural Networks	Predicted CO2 emissions from new cars, showing the potential of neural networks in understanding the impact of private transportation on emissions.
Zhe (2022)	Orthogonal Matching Pursuit Regression (OMP)	Demonstrated high accuracy in forecasting global GHG emissions per capita, showing the capability of OMP in handling complex data and making precise predictions.

2.4 Clustering of Emissions Data

Clustering techniques are pivotal for understanding patterns in greenhouse gas (GHG) emissions data across different dimensions, such as types of emissions, countries, and industries. This section explores various clustering methods and their applications in the context of GHG emissions.

2.4.1 Hierarchical Clustering by Liu et al. (2021)

Liu et al. (2021) employed hierarchical clustering to categorize countries into groups based on their GHG emissions profiles. The hierarchical clustering technique involves calculating the distance between each pair of data points (countries) using a distance metric such as Euclidean distance. This is followed by agglomerative clustering, where each country starts as a single cluster. The closest clusters are iteratively merged until all countries are grouped into a single cluster or a predetermined number of clusters is reached. The relationships between clusters are visualized using a dendrogram, which shows how clusters are merged at each step. Formula for this is:

$$D(i,j) = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \tag{2.6}$$

where $D(i,j)$ is the distance between countries i and j, and x_{ik} and x_{jk} are the values of the k-th feature (e.g., GHG emissions) for countries i and j.

2.4.2 Bibliometric Clustering by Qu et al. (2024)

Qu et al. (2024) used bibliometric analysis combined with clustering techniques to analyze research hotspots in land use carbon emissions or sinks (LUCES). The bibliometric clustering involved collecting bibliometric data from scientific databases, constructing a co-occurrence matrix to show the co-occurrence of keywords or topics in the literature, and applying clustering algorithms like k-means or hierarchical clustering to the co-occurrence matrix to identify clusters of related topics.

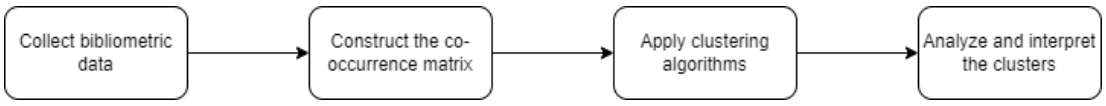

Figure 2.1 Bibliometric Clustering Flowchart

2.4.3 K-Means Clustering (Nangini et al., 2019; Cuzzocrea et al., 2015)

Nangini et al. (2019) and Cuzzocrea et al. (2015) both employed K-Means clustering to analyze greenhouse gas (GHG) emissions, albeit with different focal points and datasets. Nangini et al. (2019) focused on sector-level inventories of GHG emissions from cities, while Cuzzocrea et al. (2015) applied this technique to analyse industrial sector emissions in European countries.

Nangini et al. (2019) aimed to understand the emission patterns of various sectors within urban environments. The methodology involved normalizing the emissions data to ensure comparability across different cities and sectors. This normalization is crucial because it standardizes the data, enabling a fair comparison of emissions levels. The K-Means clustering algorithm partitions data into k clusters, where each data point is assigned to the nearest centroid. The algorithm starts by randomly selecting k initial centroids, then assigns each data point to the nearest centroid, forming clusters. The centroids are recalculated as the mean of the assigned data points, and this process iterates until the centroids no longer change significantly. The study successfully categorized cities into clusters with similar GHG emissions profiles, providing valuable insights into urban emissions patterns and highlighting differences and similarities across various urban areas.

On the other hand, Cuzzocrea et al. (2015) utilized K-Means clustering to analyse industrial sector emissions in European countries. The goal was to gain insights into emission patterns within the industrial sector, which is a significant contributor to overall GHG emissions. The process involved selecting initial centroids, assigning data points to the nearest centroid, and recalculating centroids based on the mean of the data points in each cluster. This iterative process continues until the centroids stabilize. The objective function that the K-Means algorithm aims to minimize is given by:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_j^{(i)} - \mu_i|^2 \tag{2.7}$$

where J is the objective function, k is the number of clusters, n_i is the number of points in cluster i, $x_j^{(i)}$ is the data point in cluster i, and μ_i is the centroid of cluster i. The research by Cuzzocrea et al. (2015) provided insights into industrial sector

emissions, identifying patterns and trends that can inform policy decisions and mitigation strategies.

Both studies demonstrate the effectiveness of K-Means clustering in categorizing and analysing complex GHG emissions data. By grouping similar data points together, these studies were able to identify patterns and trends that are not immediately obvious from the raw data. This clustering technique is invaluable for environmental management and policy-making, as it helps to pinpoint specific areas or sectors that require targeted interventions to reduce emissions. The insights gained from these analyses can guide efforts to mitigate GHG emissions and address the broader challenges of climate change.

. Table 2.2 Different Technique Used for Clustering

Author(s)	Technique Used	Key Findings
Liu et al. (2021)	Hierarchical Clustering	Categorized countries into groups based on GHG emissions profiles, visualized using dendrograms.
Qu et al. (2024)	Bibliometric Clustering	Analysed research hotspots in land use carbon emissions or sinks (LUCES) using bibliometric data and clustering algorithms.
Nangini et al. (2019)	K-Means Clustering	Categorized cities into clusters with similar GHG emissions profiles, highlighting urban emission patterns.
Cuzzocrea et al. (2015)	K-Means Clustering	Provided insights into industrial sector emissions in European countries, identifying patterns and trends.

2.5 GHG Emissions Visualization

Visualization techniques are essential tools in the analysis and communication of greenhouse gas (GHG) emissions data. Effective visualizations can uncover hidden patterns, trends, and relationships in complex datasets, making them more accessible

and actionable for policymakers, researchers, and the public. This section explores various methods and tools used for visualizing GHG emissions data.

2.5.1 Geographic Information System (GIS) Mapping

Geographic Information System (GIS) mapping is a powerful tool for visualizing spatial data. It allows for the representation of GHG emissions geographically, providing insights into regional variations and hotspots. For example, Ciais et al. (2023) used GIS mapping to create detailed geographic representations of CO2 emissions across different regions. These maps incorporated layers of data such as population density, industrial activity, and land use, offering a comprehensive view of the factors influencing emissions. Libraries such as geopandas, matplotlib, and folium are commonly used for GIS mapping in Python.

```
import geopandas as gpd
import matplotlib.pyplot as plt

# Load emissions data and geographic boundaries
emissions_data = pd.read_csv('emissions_data.csv')
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))

# Merge the emissions data with geographic data
geo_emissions = world.merge(emissions_data, left_on='name', right_on='country')

# Plot the map
fig, ax = plt.subplots(1, 1, figsize=(15, 10))
geo_emissions.plot(column='emissions', ax=ax, legend=True,
                   legend_kwds={'label': "GHG Emissions (MtCO2e)",
                                'orientation': "horizontal"})
plt.title('Global GHG Emissions by Country')
plt.show()
```

Figure 2.2 Pseudocode for creating a basic GIS map

2.5.2 Heat Maps

Heat maps are another popular visualization technique used to display the intensity of GHG emissions across different regions or sectors. For example, NASA (2023) utilized heat maps to illustrate global temperature anomalies and GHG emissions, making it easy to identify hotspots and areas requiring intervention. Libraries such as seaborn and matplotlib are commonly used for creating heat maps in Python.

```
import seaborn as sns
import matplotlib.pyplot as plt

# Load data
emissions_by_region = pd.read_csv('emissions_by_region.csv')

# Create heat map
plt.figure(figsize=(12, 8))
sns.heatmap(emissions_by_region.pivot('year', 'region', 'emissions'), cmap='coolwarm', annot=True)
plt.title('Heat Map of GHG Emissions by Region and Year')
plt.show()
```

Figure 2.3 Pseudocode for creating a heat map

2.5.3 Time-Series Plots

Time-series plots are crucial for tracking changes in GHG emissions over time. Mahajan and another author (2022) used time-series plots to analyze trends and anomalies in methane (CH4) concentrations over centuries, helping to understand the temporal dynamics of emissions. Libraries such as pandas, matplotlib, and plotly are widely used for creating time-series plots.

```
import pandas as pd
import matplotlib.pyplot as plt

# Load data
emissions_data = pd.read_csv('emissions_data.csv')

# Create time-series plot
plt.figure(figsize=(14, 7))
plt.plot(emissions_data['year'], emissions_data['total_emissions'], marker='o', linestyle='-')
plt.title('Time-Series of Total GHG Emissions')
plt.xlabel('Year')
plt.ylabel('Total GHG Emissions (MtCO2e)')
plt.grid(True)
plt.show()
```

Figure 2.4 Pseudocode for creating a time-series plot

2.5.4 Interactive Dashboards

Interactive dashboards provide a dynamic way to explore GHG emissions data. Ravi et al. (2020) used interactive dashboards to visualize GHG emissions from various sectors, allowing users to filter and drill down into specific details. Tools like Power BI and Tableau, as well as libraries such as dash and plotly, are commonly used to create these dashboards.

```python
import dash
import dash_core_components as dcc
import dash_html_components as html
from dash.dependencies import Input, Output
import pandas as pd

# Load data
emissions_data = pd.read_csv('emissions_data.csv')

app = dash.Dash(__name__)

app.layout = html.Div([
    dcc.Graph(id='emissions-graph'),
    dcc.Slider(
        id='year-slider',
        min=emissions_data['year'].min(),
        max=emissions_data['year'].max(),
        value=emissions_data['year'].min(),
        marks={str(year): str(year) for year in emissions_data['year'].unique()},
        step=None
    )
])

@app.callback(
    Output('emissions-graph', 'figure'),
    [Input('year-slider', 'value')]
)
def update_figure(selected_year):
    filtered_data = emissions_data[emissions_data.year == selected_year]
    return {
        'data': [{'x': filtered_data['region'], 'y': filtered_data['emissions'], 'type': 'bar'}],
        'layout': {
            'title': f'GHG Emissions for {selected_year}',
            'xaxis': {'title': 'Region'},
            'yaxis': {'title': 'GHG Emissions (MtCO2e)'}
        }
    }

if __name__ == '__main__':
    app.run_server(debug=True)
```

Figure 2.5 Example setup for an interactive dashboard in Python using Dash

2.5.5 Sankey Diagrams

Sankey diagrams are used to visualize the flow of GHG emissions from different sources to various sinks. UNEP (2023) effectively used Sankey diagrams to illustrate the flow of emissions across different sectors, highlighting the complexity of emissions sources and pathways. Libraries such as plotly and matplotlib can be used to create Sankey diagrams.

```
import plotly.graph_objects as go

# Define the Sankey diagram nodes and links
nodes = ['Sector 1', 'Sector 2', 'Sector 3', 'Atmosphere', 'Oceans', 'Land']
links = dict(
    source=[0, 1, 1, 2, 2, 3],
    target=[3, 4, 5, 4, 5, 6],
    value=[10, 5, 15, 10, 5, 20]
)

fig = go.Figure(go.Sankey(
    node=dict(
        pad=15,
        thickness=20,
        line=dict(color="black", width=0.5),
        label=nodes
    ),
    link=dict(
        source=links['source'],
        target=links['target'],
        value=links['value']
    )
))

fig.update_layout(title_text="GHG Emissions Flow", font_size=10)
fig.show()
```

Figure 2.6 Pseudocode for creating a Sankey diagram

2.5.6 Emissions Intensity Visualization

Emissions intensity visualizations, such as scatter plots and bubble charts, can display the relationship between GHG emissions and other variables, such as GDP or energy consumption. NOAA (2023) used scatter plots to show the correlation between CO2 emissions and economic indicators, highlighting the efficiency of different sectors in terms of emissions per unit of economic activity. Libraries such as matplotlib and seaborn are commonly used for these visualizations.

```
import matplotlib.pyplot as plt

# Load data
emissions_data = pd.read_csv('emissions_data.csv')

# Create scatter plot
plt.figure(figsize=(12, 8))
plt.scatter(emissions_data['gdp'], emissions_data['emissions'], c=emissions_data['region'])
plt.title('GHG Emissions vs. GDP')
plt.xlabel('GDP (in billions)')
plt.ylabel('GHG Emissions (MtCO2e)')
plt.grid(True)
plt.show()
```

Figure 2.7 Pseudocode for creating a scatter plot

2.5.7 Conclusion

Techniques like GIS mapping, heat maps, time-series plots, interactive dashboards, Sankey diagrams, and emissions intensity visualizations help stakeholders understand GHG emission sources and trends for informed policy and mitigation strategies. Authors such as Ciais et al. (2023) and Mahajan (2022) have effectively used these tools to present and analyse emissions data.

. Table 2.3 Summary of Visualization Methods

Method	Description	Libraries	Use Cases
GIS Mapping	Visualizes geographic data, showing regional variations and hotspots	geopandas, folium	Mapping global GHG emissions by country (Ciais et al., 2023)
Heat Maps	Uses colour gradients to show intensity of emissions across regions/sectors	seaborn, matplotlib	Identifying emission hotspots (NASA, 2023)
Time-Series Plots	Tracks changes in emissions over time, showing trends and anomalies	pandas, matplotlib, plotly	Analyzing temporal dynamics of emissions (Mahajan and another author, 2022)
Interactive Dashboards	Integrates various visualizations for dynamic data exploration	dash, plotly, Power BI	Exploring emissions data interactively (Ravi et al., 2020)
Sankey Diagrams	Visualizes the flow of emissions from sources to sinks	plotly, matplotlib	Illustrating emissions flow across sectors (UNEP, 2023)
Emissions Intensity Visualization	Displays relationships between emissions and other variables	matplotlib, seaborn	Showing correlation between CO2 emissions and GDP (NOAA, 2023)

2.6 Performance Measurement

Assessing the accuracy and reliability of models that forecast greenhouse gas (GHG) emissions is essential. By utilizing the right metrics and methodologies, one can determine how well these models identify patterns within the data and how accurately they can make predictions. This section outlines various methods and metrics employed in the performance evaluation of GHG emissions models.

Root Mean Square Error (RMSE) is a widely used metric for assessing the accuracy of predictive models. It measures the average magnitude of the errors between predicted and observed values, providing an indication of the model's overall accuracy. RMSE is particularly useful in GHG emissions modeling as it penalizes larger errors more heavily, making it sensitive to outliers. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.8}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and n is the number of observations. For instance, in a study by Mahajan et al. (2022), RMSE was used to evaluate the performance of their predictive models for CH4 concentrations. The models achieved an RMSE of 0.95, indicating high accuracy in the predictions.

Mean Absolute Error (MAE) measures the average magnitude of errors in a set of predictions, without considering their direction. It provides a straightforward interpretation of the average error made by the model and is less sensitive to outliers compared to RMSE. The formula for MAE is:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.9}$$

In their analysis of GHG emissions, Wang et al. (2022) utilized MAE to assess the accuracy of their machine learning models. The models showed an MAE of 1.2 MtCO2e, reflecting the average deviation of predictions from actual values.

The Coefficient of Determination (R^2) measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

It indicates how well the model explains the variability of the outcome and ranges from 0 to 1, with higher values indicating better model performance. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.10}$$

where $\bar{y}$ is the mean of the observed data. In their study, Julian et al. (2023) reported an R^2 of 0.87 for their supervised machine learning model predicting CO2 emissions from new cars. This high R^2 value indicates that the model explains 87% of the variance in the emissions data.

Cross-validation is a technique used to assess the generalizability of a model. By partitioning the data into training and testing sets multiple times, cross-validation ensures that the model's performance is not dependent on a particular split of the data. K-fold cross-validation is commonly used, where the data is divided into k subsets, and the model is trained and tested k times, each time using a different subset as the test set. Ravi et al. (2020) applied 10-fold cross-validation in their interactive dashboard model to ensure robustness and prevent overfitting. The average performance across the folds provided a reliable estimate of the model's accuracy.

Precision and recall are metrics used in classification models to evaluate the quality of predictions. Precision measures the proportion of true positive predictions out of all positive predictions, while recall measures the proportion of true positive predictions out of all actual positives. The formulas for precision and recall are:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2.11}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2.12}$$

In their study on GHG emissions, Sheng et al. (2022) used precision and recall to evaluate the performance of their classification model in identifying high-emission regions. The model achieved a precision of 0.82 and a recall of 0.78, indicating a good balance between accuracy and coverage.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction to the Framework

In this research, a comprehensive and structured framework is essential for systematically investigating the complex dynamics of global greenhouse gas (GHG) emissions. The framework adopted integrates various methodologies to ensure a thorough analysis, accurate prediction, and continuous improvement of the models and strategies developed. The framework consists of six key phases, each designed to address specific aspects of the research process, from initial project setup to ongoing refinement and enhancement of the results, Such as:

I. Project Initialization

II. Data Exploration

III. Data Mining

IV. Prediction Models

V. Performance Evaluation and Knowledge Sharing

VI. Continuous Improvement

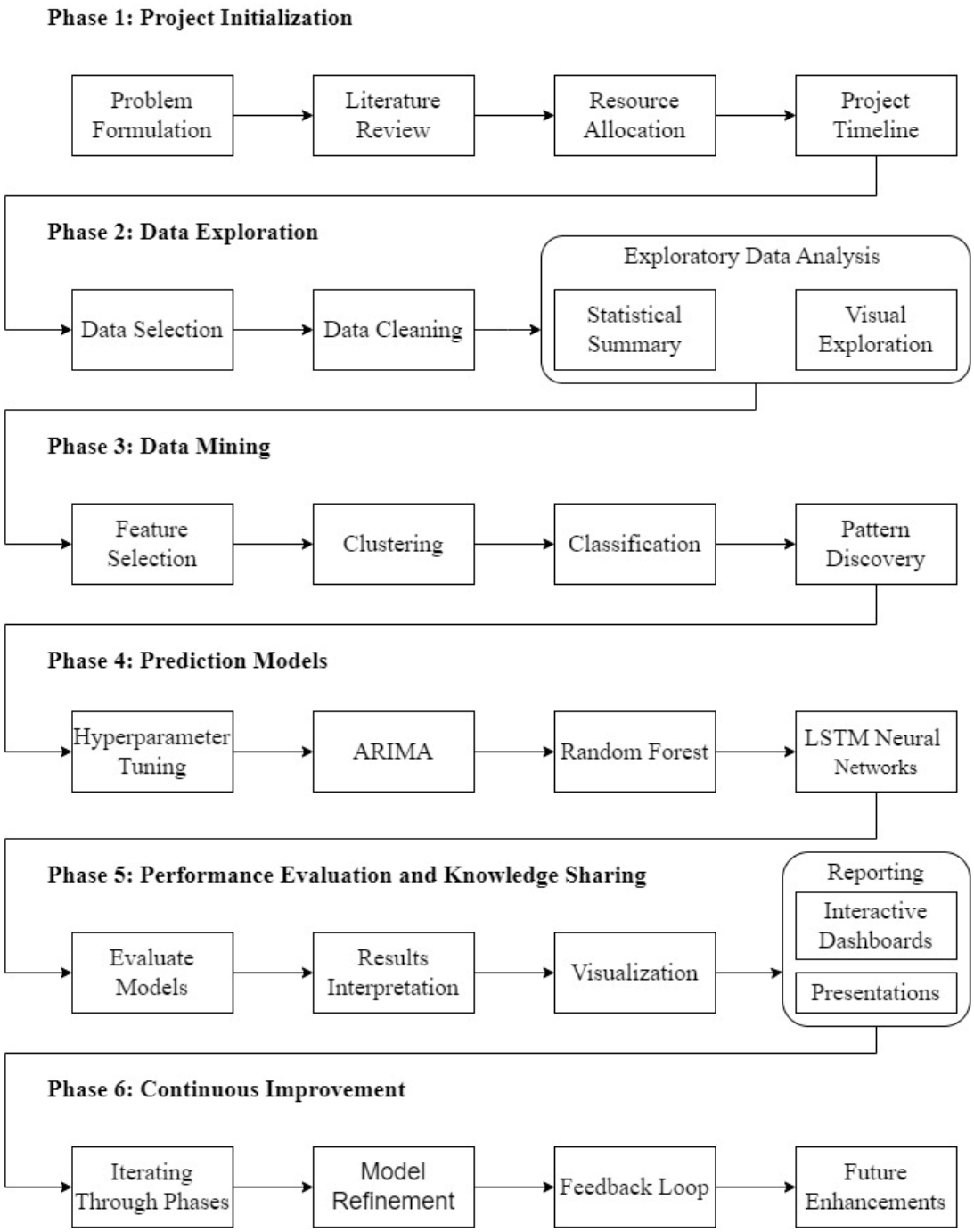

Figure 3.1 Six Phases Research Framework

3.2 Project Initialization

The project initialization phase is fundamental to setting the stage for this research. This phase encompasses the formulation of the research problem, conducting a comprehensive literature review, allocating necessary resources, and establishing a

detailed project timeline. These steps are crucial for ensuring a structured approach to the study, guiding the research from its inception through to its completion.

3.2.1 Problem Formulation

The escalation of global greenhouse gas (GHG) emissions is a major contributor to climate change, impacting environmental health, economic stability, and societal well-being. Despite international efforts such as the Paris Agreement, GHG levels continue to rise, driven by industrial activities, transportation, energy production, and agricultural practices. The complexity of these sources and the lack of a comprehensive predictive model that incorporates various influential factors hinder accurate projections of future emissions. This research aims to address the critical issue of accurately forecasting global GHG emissions. The objectives of this study are to analyse seasonal variations, long-term trends, and cyclical patterns in GHG emissions data from 1970 to 2019; assess the impact of different sectors such as agriculture, industry, transportation, and energy on global GHG emissions; investigate how GHG emissions vary across different regions and identify factors contributing to these differences through data analysis; create robust predictive models for GHG emissions; and develop data-driven strategies for mitigating GHG emissions based on comprehensive analysis and predictive modelling.

3.2.2 Literature Review

A thorough literature review was conducted to understand the current state of research on GHG emissions and predictive modeling techniques. Studies such as those by Mahajan and Jain (2022) utilized exploratory data analysis and time-series forecasting to understand the causes of global warming and predict methane (CH4) concentrations with high accuracy. Wang et al. (2022) highlighted the potential of machine learning models in forecasting global GHG emissions per capita, showcasing the effectiveness of these models in capturing complex patterns in emissions data. Additionally, Ciais et al. (2023) provided insights into significant emissions reductions during the COVID-19 pandemic using near-real-time monitoring, which underscores the impact of global events on emissions. By integrating these diverse perspectives,

the research identified gaps in existing models, particularly the need for incorporating a broader range of factors and higher temporal resolution in emissions data.

Further studies, such as those by Birol et al. (2023) and Ritchie et al. (2023), have provided comprehensive analyses and forecasts of CO2 emissions, emphasizing the importance of using robust datasets like those from the International Energy Agency (IEA) and Our World in Data. These sources offer detailed and reliable emissions data, which are crucial for developing accurate predictive models. The literature review concluded that there is a significant opportunity to enhance current forecasting models by integrating advanced data science techniques and a wider array of data sources, thereby improving the accuracy and reliability of GHG emissions predictions.

3.2.3 Resource Allocation

The next step involves allocating the necessary resources for the research. This includes identifying and securing the datasets needed for analysis, such as the IMF's Quarterly Greenhouse Gas (GHG) Air Emissions Accounts dataset. This dataset was selected due to its comprehensive coverage of GHG emissions on a quarterly basis, providing high temporal resolution and including various sectors and countries. Additionally, it is crucial to ensure access to advanced data analysis tools and software, such as Python libraries (Pandas, NumPy, Matplotlib, Seaborn) and machine learning frameworks (Scikit-learn, TensorFlow, Keras). Having the right tools and resources is essential for the successful execution of the project. As an individual researcher, ensuring efficient utilization of these resources will be key to managing the project effectively and achieving the research objectives.

3.2.4 Project Timeline

Establishing a detailed project timeline is critical for ensuring that the research progresses in a structured and timely manner. The timeline for my master project begins in weeks 1 to 2 with a focus on problem formulation to clearly define the research question and objectives. This is followed by a comprehensive literature review in weeks 3 to 5 to gather relevant studies and identify gaps in existing research.

Data collection takes place in weeks 6 to 7, sourcing necessary data from repositories such as the IMF's Quarterly Greenhouse Gas (GHG) Air Emissions Accounts dataset. Preliminary data analysis is conducted in weeks 8 to 9 to understand the basic characteristics and structure of the data. In weeks 10 to 11, data wrangling is performed to clean and preprocess the data, addressing any missing values or inconsistencies. Exploratory data analysis (EDA) occurs in weeks 12 to 13 to uncover patterns, trends, and insights within the data. Finally, weeks 14 to 15 are dedicated to preparing the draft of the Project 1 Report, incorporating the findings from the EDA.

Throughout this period, I will have regular and structured meetings with my supervisor, Prof. Madya Dr. Mohd Shahizan Othman. His expertise and guidance will be instrumental in refining my research approach and ensuring that I meet the project deadlines. Regular consultations will help in addressing any challenges promptly and keeping the project on track. The oral presentation and proposal defense will provide a platform to present my findings and receive feedback in final week.

3.3 Data Exploration

The data exploration phase is vital for establishing a foundational understanding of the dataset being analyzed. This phase involves the meticulous selection, cleaning, and preliminary examination of the data to uncover its inherent patterns and structures. By ensuring the data is well-prepared and thoroughly understood, this phase sets the stage for effective data mining and modeling efforts in subsequent stages.

3.3.1 Data Selection

For this research, the Quarterly Greenhouse Gas (GHG) Air Emissions Accounts dataset from the International Monetary Fund (IMF) was chosen. This dataset was selected because it provides comprehensive and detailed coverage of GHG emissions on a quarterly basis, offering high temporal resolution and encompassing various sectors and countries. The extensive nature of the IMF dataset makes it ideal

for analyzing emissions trends and identifying significant contributing factors. Its multi-decade span allows for the examination of long-term trends and changes.

Table 3.1 Selected Dataset Attributes

Attribute	Description
ObjectId	Unique identifier for each record
Country	Name of the country or group of countries
ISO2	Two-letter country code
ISO3	Three-letter country code
Indicator	Type of GHG emission measured
Unit	Unit of measurement (e.g., million metric tons of CO2 equivalent)
Source	Origin of the data
CTS_Code	Code representing the data series
CTS_Name	Name of the data series
CTS_Full_Descriptor	Full descriptor of the data series
Industry	Economic sector the data pertains to
Gas_Type	Specific type of greenhouse gas measured
Seasonal_Adjustment	Indicates whether the data is seasonally adjusted
Scale	Scale of the measurement units
F20XXQY	Emissions data for the Yth quarter of the year 20XX

This dataset is particularly valuable due to its comprehensive scope and high-quality data, making it a robust foundation for analysing GHG emissions and developing predictive models. The dataset's memory size is approximately 15 MB, ensuring it is manageable for detailed analysis while offering extensive data for robust insights.

By leveraging this dataset, the research aims to uncover meaningful patterns, trends, and relationships within GHG emissions data, contributing to the development of effective mitigation strategies and informed policy decisions.

```
import pandas as pd
df = pd.read_csv('/Quarterly_Greenhouse_Gas_(GHG)_Air_Emissions_Accounts.csv')
df
```

	ObjectId	Country	ISO2	ISO3	Indicator	Unit	Source	CTS_Code	CTS_Name	CTS_F
0	1	Advanced Economies	NaN	AETMP	Quarterly greenhouse gas (GHG) air emissions a...	Million metric tons of CO2 equivalent	Organisation for Economic Co-operation and Dev...	ECNGA	Greenhouse Gas Emissions (GHG); Air Emissions ...	Envirc Chanç
1	2	Advanced Economies	NaN	AETMP	Quarterly greenhouse gas (GHG) air emissions a...	Million metric tons of CO2 equivalent	Organisation for Economic Co-operation and Dev...	ECNGA	Greenhouse Gas Emissions (GHG); Air Emissions ...	Envirc Chanç
2	3	Advanced Economies	NaN	AETMP	Quarterly greenhouse gas (GHG) air emissions a...	Million metric tons of CO2 equivalent	Organisation for Economic Co-operation and Dev...	ECNGA	Greenhouse Gas Emissions (GHG); Air Emissions ...	Envirc Chanç

Figure 3.2 Reading Selected Dataset

From Figure 3.2, we can see the initial steps of our data selection process, where we imported the pandas library and read the dataset file into a DataFrame. This involved using the pd.read_csv() function to load the csv file. The dataset was then displayed to provide an overview of its structure and contents. These initial steps are crucial as they set the stage for further data cleaning and analysis, ensuring that we have successfully imported and visualized the dataset for subsequent processing.

3.3.2 Data Cleaning

The data cleaning process involves several critical steps to ensure the dataset is ready for analysis. These steps include checking the data's structure and quality, handling missing values, removing duplicates, and standardizing the data for consistency. First, we examine the structure and basic information of the dataset using the df.info() function. This step provides a summary of the data, including the number of entries, columns, data types, and non-null counts for each column..

The initial inspection of the dataset reveals that it contains 2,372 entries and 68 columns. Most of the columns are completely filled, but the ISO2 column is entirely empty, indicating it will likely be removed in subsequent steps. The data types are also provided, which helps in identifying any necessary type conversions. This initial

review, as shown in Figure 3.3, highlights the dataset's structure and any immediate issues, such as missing values or incorrect data types.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2372 entries, 0 to 2371
Data columns (total 68 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ObjectId              2372 non-null   int64
 1   Country               2372 non-null   object
 2   ISO2                  0 non-null      float64
 3   ISO3                  2372 non-null   object
 4   Indicator             2372 non-null   object
 5   Unit                  2372 non-null   object
 6   Source                2372 non-null   object
 7   CTS_Code              2372 non-null   object
 8   CTS_Name              2372 non-null   object
 9   CTS_Full_Descriptor   2372 non-null   object
 10  Industry              2372 non-null   object
 11  Gas_Type              2372 non-null   object
 12  Seasonal_Adjustment   2372 non-null   object
 13  Scale                 2372 non-null   object
 14  F2010Q1               2372 non-null   float64
```

Figure 3.3 Data Information of Research Dataset

Next, we proceed to remove duplicate records from the dataset using df.drop_duplicates(). Removing duplicates ensures that each record is unique, preventing any distortion in the analysis. In this dataset, no duplicates were found, so the number of entries remains the same. The next step is to drop columns with a significant number of missing values, such as the ISO2 column.

```
df.drop(columns=['ISO2'], inplace=True)
df
```

	ObjectId	Country	ISO3	Indicator	Unit	Source	CTS_Code	CTS_I
0	1	Advanced Economies	AETMP	Quarterly greenhouse gas (GHG) air emissions a...	Million metric tons of CO2 equivalent	Organisation for Economic Co-operation and Dev...	ECNGA	Greenh Emis (GHG Emis

Figure 3.4 ISO2 Column Drop

The ISO2 column was dropped as it had no entries. This step helps in maintaining the quality of the dataset by removing irrelevant data. Next, we check and correct the data types if necessary. The data types are mostly accurate, with numerical data stored as float64 and categorical data as objects. No major corrections are needed here, as shown in Figure 3.5.

```
data_types = df.dtypes
print(data_types[0:15])
```

```
ObjectId                 int64
Country                 object
ISO3                    object
Indicator               object
Unit                    object
Source                  object
CTS_Code                object
CTS_Name                object
CTS_Full_Descriptor     object
Industry                object
Gas_Type                object
Seasonal_Adjustment     object
Scale                   object
F2010Q1                float64
F2010Q2                float64
dtype: object
```

Figure 3.5 Data Types Check

Next, we identify and remove unnecessary columns to streamline the dataset. To achieve this, we first check the number of unique values in each column using df.nunique(), as shown in Figure 3.6. This step helps in identifying columns with only one unique value, which do not contribute to the analysis and can be removed. From Figure 3.6, we observe that several columns, such as Indicator, Unit, Source, CTS_Code, CTS_Name, CTS_Full_Descriptor, and Scale, have only one unique value. These columns do not provide any additional information and can be safely removed. We then proceed to drop these columns using the df.drop() function, as illustrated in Figure 3.7. By removing columns with less than one unique count, we focus the dataset on relevant information, making it more manageable for analysis. This process ensures that the dataset is streamlined and free from redundant columns, thereby enhancing the efficiency of subsequent data analysis and modelling efforts.

```
unique_value_counts = df.nunique()
unique_value_counts[0:15]
```

```
ObjectId              2372
Country                 25
ISO3                    25
Indicator                1
Unit                     1
Source                   1
CTS_Code                 1
CTS_Name                 1
CTS_Full_Descriptor      1
Industry                10
Gas_Type                 5
Seasonal_Adjustment      2
Scale                    1
F2010Q1               2333
F2010Q2               2333
dtype: int64
```

Figure 3.6 Checking Unnecessary columns

```
columns_to_drop = unique_value_counts[unique_value_counts == 1].index.tolist()
df.drop(columns=columns_to_drop, inplace=True)
df
```

	ObjectId	Country	ISO3	Industry	Gas_Type	Seasonal_Adjustment	F2010Q1	F201(
0	1	Advanced Economies	AETMP	Agriculture, Forestry and Fishing	Carbon dioxide	Not Seasonally Adjusted	43.908353	48.159
1	2	Advanced Economies	AETMP	Agriculture, Forestry and Fishing	Carbon dioxide	Seasonally Adjusted	47.679026	48.234
		Advanced		Agriculture,	Fluorinated	Not Seasonally		

Figure 3.7 Dropping Column with less than 1 Unique Count

To ensure the dataset's completeness, we check for missing values in each column using the df.isna().sum() function, as shown in Figure 3.7. This step helps identify any gaps in the data that need to be addressed before proceeding with the analysis. we can see that there are no missing values in any of the columns, indicating that the dataset is complete and ready for further analysis. This verification step is crucial as it ensures that our analysis will not be biased by incomplete data, thereby enhancing the reliability of our results.

```
missing_values = df.isna().sum()
print(missing_values)
```

```
ObjectId                0
Country                 0
ISO3                    0
Industry                0
Gas_Type                0
Seasonal_Adjustment     0
F2010Q1                 0
F2010Q2                 0
```

Figure 3.8 Checking Missing Values

```
df = df.rename(columns=lambda x: x.replace('F', '') if x.startswith('F') else x)
df
```

	ObjectId	Country	ISO3	Industry	Gas_Type	Seasonal_Adjustment	2010Q1	2010Q2	2010Q3
0	1	Advanced Economies	AETMP	Agriculture, Forestry and Fishing	Carbon dioxide	Not Seasonally Adjusted	43.908353	48.159155	51.673939
1	2	Advanced Economies	AETMP	Agriculture, Forestry and Fishing	Carbon dioxide	Seasonally Adjusted	47.679026	48.234737	49.320131

Figure 3.9 Renaming Quarters Column

To improve clarity and consistency, we rename the columns, particularly focusing on the quarterly data columns. This step involves using a lambda function within the df.rename() method to remove the prefix 'F' from the quarterly column names, making them easier to read and understand. By implementing this renaming process, we ensure that the column names are straightforward and reflect the data they contain without unnecessary prefixes. This adjustment helps in maintaining a clean and easily interpretable dataset, which is crucial for subsequent analysis.

To identify the outliers, we used the Z-score method. The Z-score measures how many standard deviations a data point is from the mean. A common threshold for identifying outliers is a Z-score greater than 3 or less than -3. This method was implemented by first calculating the Z-scores for the numeric columns in the dataset, as shown in Figure 3.10.

From the Z-scores, we determined the number of outliers per row. Any row with a Z-score exceeding the threshold in any of its numeric columns was marked as an outlier. The outliers_df DataFrame was created to store these rows, containing the

columns ObjectId and Num_Outliers which indicate the number of outliers found in each row.

To handle these outliers, we chose to impute them with the mean value of the same quarter across different years. This approach helps maintain the integrity of the dataset by reducing the impact of extreme values without simply removing data points. The imputation process is illustrated in Figure 3.11.

```
from scipy import stats
numeric_columns = df.columns[6:]
z_scores = stats.zscore(df[numeric_columns], axis=1)
threshold = 3
num_outliers_per_row = (z_scores > threshold).sum(axis=1)
df['Num_Outliers'] = num_outliers_per_row
outliers_df = df[df['Num_Outliers'] > 0]
print(outliers_df[['ObjectId', 'Num_Outliers']])
```

```
      ObjectId  Num_Outliers
17          18             1
87          88             1
94          95             1
132        133             1
133        134             1
...        ...           ...
2027      2042             1
2167      2182             1
2169      2184             1
2211      2226             1
2273      2288             1

[87 rows x 2 columns]
```

Figure 3.10 Finding Outliers

```
outlier_indices = outliers_df['ObjectId'].values
for index in outlier_indices:
    for column in df.columns:
        if column.endswith('Q1') or column.endswith('Q2') or column.endswith('Q3') or column.endswith('Q4'):
            quarter = column[-2:]  # Extracting the quarter part from the column name
            same_quarter_columns = [col for col in df.columns if col.endswith(quarter) and col != column]
            same_quarter_mean = df[same_quarter_columns].mean(axis=1).iloc[index]
            df.at[index, column] = same_quarter_mean
```

Figure 3.11 Handling Outliers

In this process, we first extracted the indices of the outliers from the outliers_df DataFrame. For each outlier index, we iterated through each column to check if it represented a quarter (ending with 'Q1', 'Q2', 'Q3', or 'Q4'). For columns identified as

quarters, we calculated the mean of the same quarter across different years, excluding the current outlier value. The outlier value was then replaced with this mean value. This method, as shown in Figure 3.11, ensures that the data remains consistent and representative of typical values for each quarter. Through these steps, we effectively handled the outliers in our dataset, ensuring that the data is clean and reliable for further analysis. After handling the outliers, we removed the Num_Outliers column from the dataset to clean up any temporary columns used during the process. This was done using the code df = df.drop('Num_Outliers', axis=1). Removing this column ensures that the dataset remains tidy and only contains relevant information for further analysis. Finally, we saved the cleaned dataset to a new CSV file using the code df.to_csv('cleaned_quarterly_ghg_emissions.csv', index=False). This step preserves the cleaned data for subsequent use and analysis.

3.3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) encompasses both statistical summaries and visual exploration, offering a detailed understanding of the data's characteristics and interrelations. The goal of EDA is to gain deeper insights into the dataset, identify any anomalies, validate assumptions, and develop hypotheses. Through EDA, we can uncover critical patterns, relationships, and insights that are essential for the subsequent phases of data mining and predictive modeling. Combining statistical and visual techniques ensures a comprehensive examination of the data from various angles, ensuring no crucial details are overlooked.

Our EDA process involved a thorough initial analysis and the use of visualization techniques to better understand the dataset. Histograms were employed to analyze the distribution of individual variables, while box plots helped in identifying outliers and understanding the data's spread. Scatter plots were used to investigate relationships between pairs of variables, and heatmaps provided a visual representation of correlations among variables. These visual tools, created using libraries such as Matplotlib, Seaborn, and Plotly, offered valuable insights into the data. Alongside visual exploration, we generated a statistical summary, including descriptive statistics

like mean, median, and standard deviation, as well as correlation analysis using Pearson and Spearman correlation coefficients. Trend analysis was conducted to identify temporal patterns within the data. These analyses, performed with Pandas and NumPy, enabled us to extract meaningful insights and highlight key trends and relationships within the dataset.

3.3.3.1 Statistical Summary

The statistical summary involves generating descriptive statistics and conducting correlation analysis to understand the dataset's properties. The descriptive statistics provide a comprehensive summary of the dataset, offering insights into various aspects such as the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values for each variable. As depicted in Figure 3.12, the dataset comprises 2,372 entries for each quarter from 2010 to 2023. For the quarter-specific columns like 2010Q1, the mean value is around 198.92 with a standard deviation of 749.75, reflecting considerable variability in greenhouse gas emissions across different data points. The minimum value for 2010Q1 is -0.012365, indicating some negative values that might be anomalies or data entry errors. The 25th percentile value is 0.800094, the median is 10.263661, and the 75th percentile is 88.293932, showing the interquartile range and the spread of the data. The maximum value for this quarter is 12108.179190.

Similar patterns are observed across other quarters, such as 2010Q2, 2010Q3, 2010Q4, and 2011Q1, with mean values and standard deviations indicating variations in emissions. The minimum and maximum values, along with the quartiles, provide a detailed view of the data distribution, revealing the central tendency of the dataset. Overall, the descriptive statistics underscore the extensive variability in the dataset, emphasizing the importance of thorough data cleaning and normalization to ensure accurate analysis. The comprehensive statistical summary aids in identifying key trends, anomalies, and patterns, setting the foundation for more detailed exploratory data analysis.

```
descriptive_stats = df.describe()
descriptive_stats
```

	ObjectId	2010Q1	2010Q2	2010Q3	2010Q4	2011Q1
count	2372.000000	2372.000000	2.372000e+03	2372.000000	2372.000000	2.372000e+03
mean	1199.549747	198.918807	1.922395e+02	193.705582	203.224204	2.090696e+02
std	686.423558	749.746453	7.301587e+02	738.074724	771.595443	7.940926e+02
min	1.000000	-0.012365	-1.780000e-15	-0.012141	-0.028861	-1.720000e-15
25%	607.750000	0.800094	8.192225e-01	0.842045	0.844306	8.425395e-01
50%	1200.500000	10.263661	9.787214e+00	9.817681	10.618808	1.047325e+01
75%	1793.250000	88.293932	8.795017e+01	86.314611	91.628301	9.327457e+01
max	2386.000000	12108.179190	1.157531e+04	11701.016270	12104.069020	1.271917e+04

8 rows × 55 columns

Figure 3.12 Descriptive Statistics

The correlation analysis provides insights into the relationships between different variables in the dataset. The Pearson correlation, as shown in Figure 3.13, measures the linear relationship between pairs of variables. A correlation coefficient close to 1 indicates a strong positive linear relationship, while a coefficient close to -1 indicates a strong negative linear relationship. Values close to 0 suggest no linear relationship.

From the results, we observed that the correlation between different quarters is very high, often close to 1, indicating that the greenhouse gas emissions data for different quarters are strongly positively correlated. For instance, the correlation between 2010Q1 and 2010Q2 is 0.996609, and between 2010Q1 and 2010Q3 is 0.995847, reflecting consistent trends in emissions across these periods. This high correlation suggests that emissions data tend to follow a similar pattern across consecutive quarters, likely due to underlying seasonal trends or consistent reporting methods.

Similarly, the Spearman correlation, also depicted in Figure 3.13, measures the monotonic relationship between variables, which can capture both linear and non-linear associations. The Spearman correlation coefficients are also high, indicating strong monotonic relationships between the quarterly data points. For example, the

correlation between 2010Q1 and 2010Q2 is 0.996347, and between 2010Q1 and 2010Q3 is 0.992872, showing a strong rank-order relationship.

These high correlations underscore the consistency and reliability of the emissions data across different quarters. They also highlight the importance of accounting for temporal dependencies in any further analysis or predictive modeling. By understanding these relationships, we can better interpret the trends and patterns in the dataset, leading to more accurate and insightful conclusions about greenhouse gas emissions.

```
numeric_df = df.select_dtypes(include=[np.number])

pearson_corr = numeric_df.corr(method='pearson')
spearman_corr = numeric_df.corr(method='spearman')

print("Pearson Correlation:\n", pearson_corr)
print("Spearman Correlation:\n", spearman_corr)
```

```
Pearson Correlation:
           ObjectId    2010Q1    2010Q2    2010Q3    2010Q4    2011Q1  \
ObjectId  1.000000 -0.026097 -0.025062 -0.025315 -0.025384 -0.026123
2010Q1   -0.026097  1.000000  0.996609  0.995847  0.999261  0.999606
2010Q2   -0.025062  0.996609  1.000000  0.999797  0.998710  0.996955
2010Q3   -0.025315  0.995847  0.999797  1.000000  0.998251  0.996186
2010Q4   -0.025384  0.999261  0.998710  0.998251  1.000000  0.999432
2011Q1   -0.026123  0.999606  0.996955  0.996186  0.999432  1.000000
2011Q2   -0.025222  0.995852  0.999705  0.999582  0.998424  0.996776
2011Q3   -0.025554  0.995319  0.999495  0.999734  0.998083  0.996202
```

Figure 3.13 Correlation Analysis

To analyze the temporal patterns and understand changes over time, we performed trend analysis by aggregating the greenhouse gas emissions data on a yearly basis. The following code reshapes the data, extracts the year from each quarter, calculates the mean emissions for each year, and plots the resulting trend. This comprehensive approach helps in identifying long-term trends and variations in the dataset. The code, as shown in Figure 3.14, first melts the DataFrame to convert quarterly columns into rows, making it easier to manipulate. It then extracts the year from each quarter and groups the data by year to calculate the mean emissions. Finally, the trend of mean emissions over the years is plotted.

```
df_melted = df.melt(id_vars=['ObjectId', 'Country', 'ISO3', 'Industry', 'Gas_Type', 'Seasonal_Adjustment'],
                    var_name='Quarter', value_name='Emissions')

df_melted['Year'] = df_melted['Quarter'].apply(lambda x: int(x[:4]))
yearly_data = df_melted.groupby('Year')['Emissions'].mean().reset_index()

# Plot the trend of mean emissions over the years
plt.figure(figsize=(15, 6))
plt.plot(yearly_data['Year'], yearly_data['Emissions'], marker='o')
plt.title('Trend Analysis of Mean Emissions Over Years')
plt.xlabel('Year')
plt.ylabel('Mean Emissions')
plt.grid(True)
plt.show()
```

Figure 3.14 Code for Trend Analysis

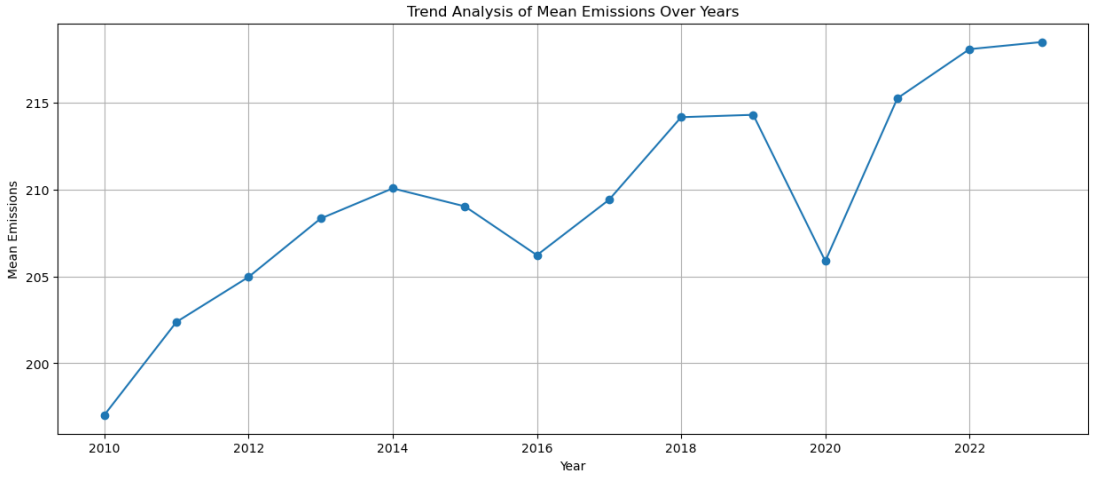

Figure 3.15 Trend Analysis

The resulting graph, as depicted in Figure 3.15, shows the trend of mean greenhouse gas emissions from 2010 to 2023. This visual representation helps in understanding how emissions have changed over time. From the graph, we can observe a general upward trend in mean emissions over the years. However, there is a noticeable dip around 2020, which corresponds to the COVID-19 pandemic period. This dip likely reflects the global reduction in industrial activity and transportation during the pandemic, leading to a temporary decrease in emissions. Following this period, the emissions trend resumes its upward trajectory, indicating a recovery in economic activities and corresponding emissions levels.

This trend analysis provides valuable insights into the temporal dynamics of greenhouse gas emissions, highlighting significant patterns and deviations that are crucial for understanding the overall impact and for making informed decisions in environmental policy and planning.

3.3.3.2 Visual Exploration

Visual exploration is important for understanding and interpreting the dataset. It involves using various graphical techniques to identify patterns, relationships, and anomalies in the data. By leveraging visual tools such as histograms, box plots, scatter plots, and heatmaps, we can gain valuable insights that might not be apparent from raw data alone. Additionally, other visual tools available in pandas and matplotlib, such as line charts, bar charts, and stack plots, can be used to answer different types of research questions. These tools allow us to explore trends, distributions, correlations, and compositions within the dataset, providing a comprehensive view of the data through different graphical representations.

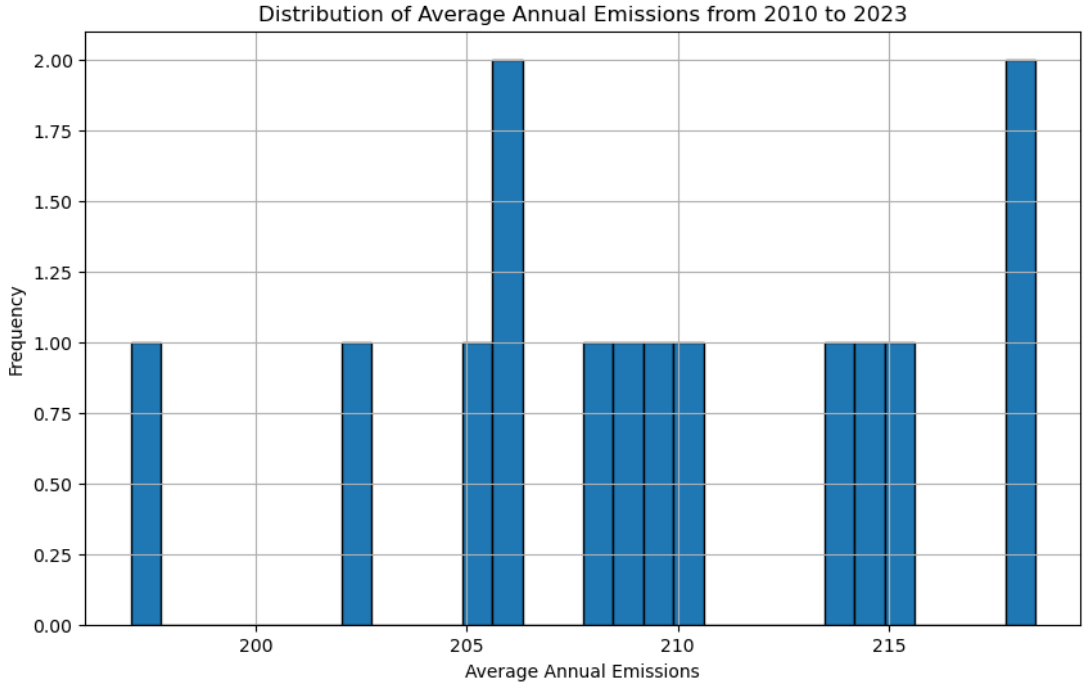

Figure 3.16 Distribution of Average Annual Emissions from 2010 to 2023

The histogram depicted in Figure 3.16 illustrates the distribution of average annual emissions from 2010 to 2023. This graph helps in understanding how emissions are spread over the years, indicating the frequency of different emission levels. It is clear from the histogram that there are peaks around the 205 and 215 levels, suggesting these values occur more frequently within the dataset.

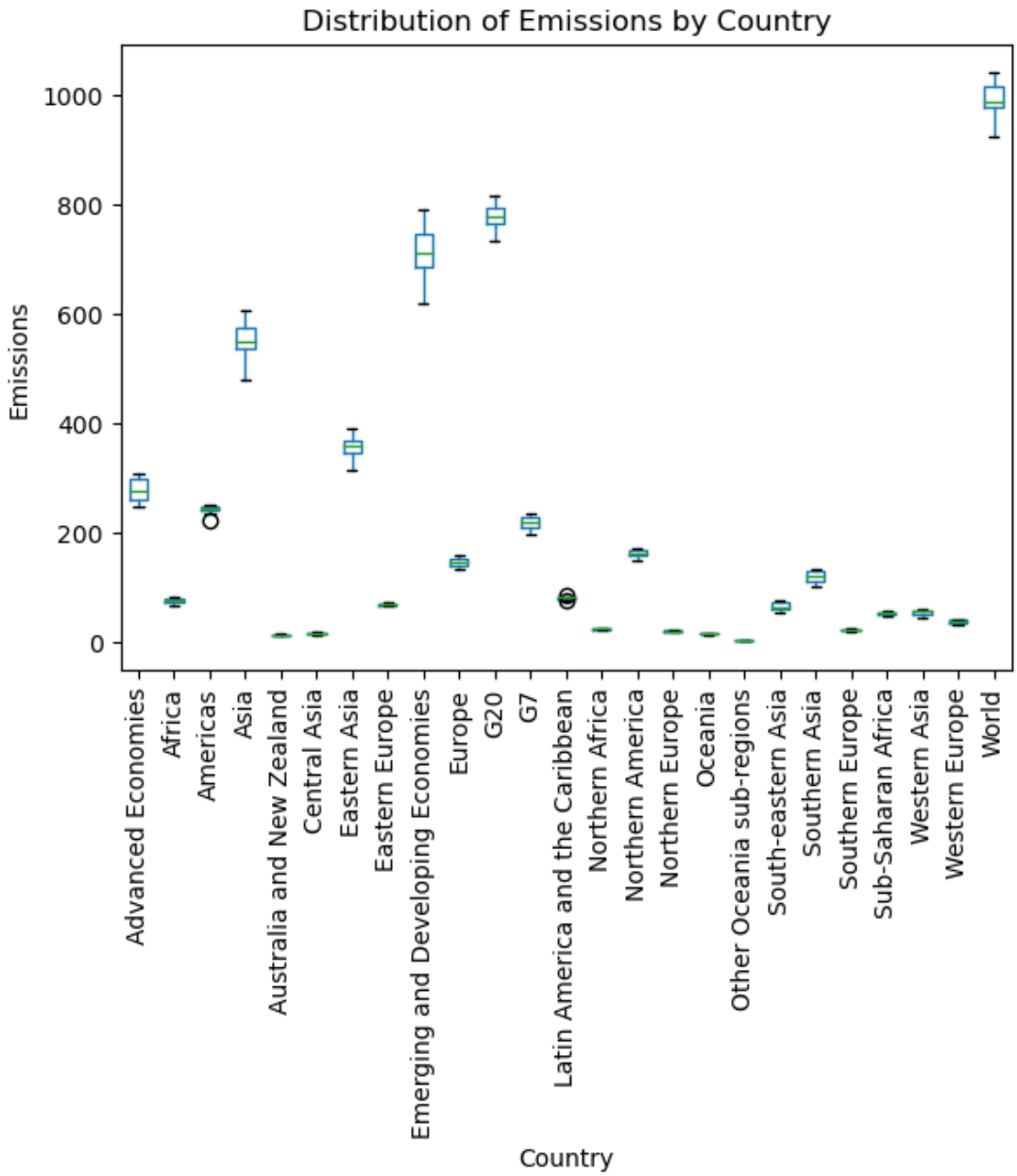

Figure 3.17 Distribution of Emissions by Country

In Figure 3.17, a box plot is utilized to display the distribution of emissions across different countries. This plot is beneficial in identifying the variability in emissions within and between countries. Notably, Advanced Economies, Eastern Asia, and Western Europe show higher median emission values, with some countries exhibiting significant variability. This information can be beneficial for policymakers focusing on emission reduction strategies in specific regions.

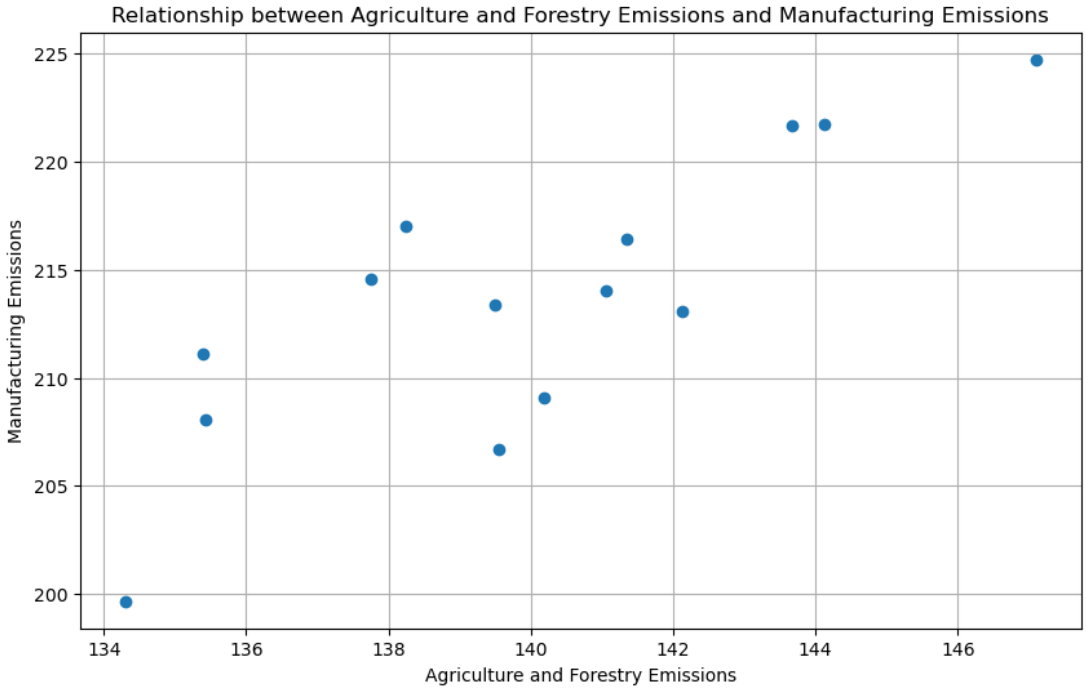

Figure 3.18 Relationship between Agriculture and Forestry and Manufacturing

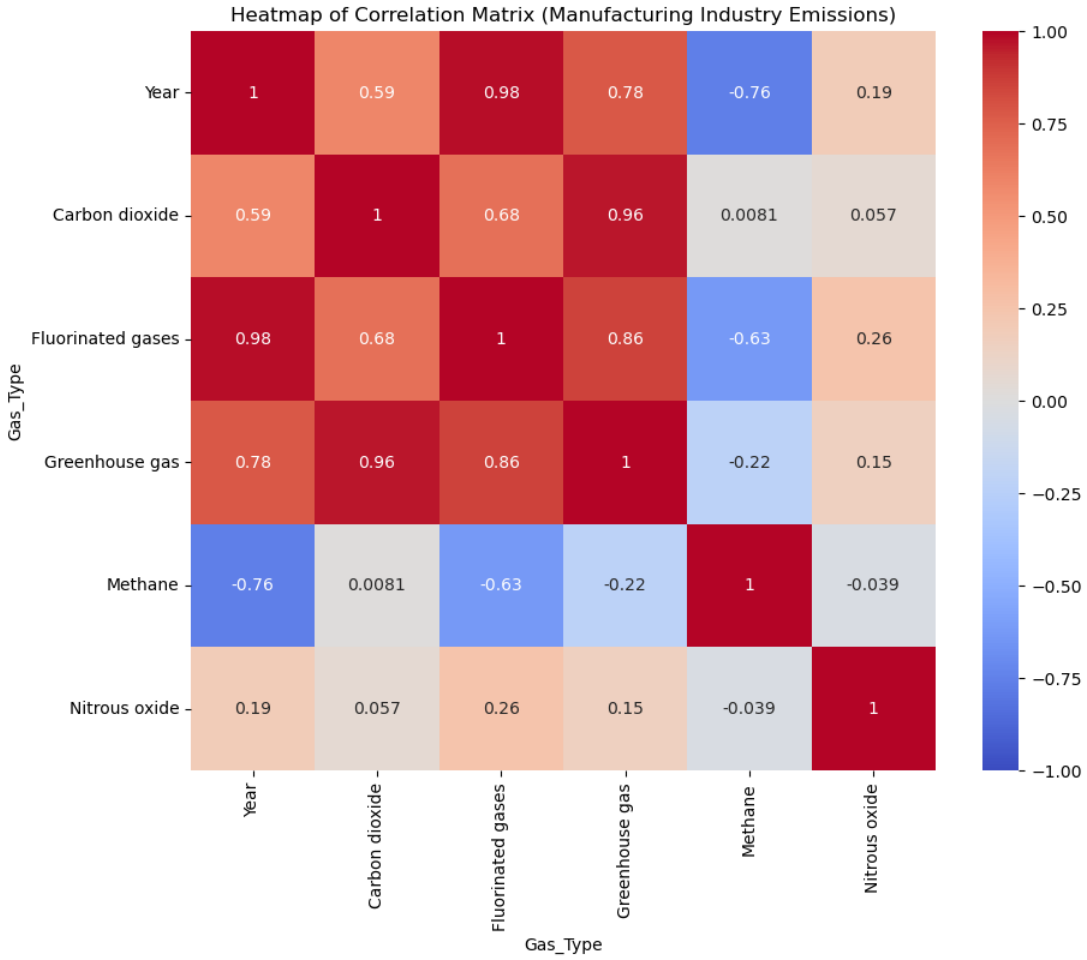

Figure 3.19 Heatmap of Correlation Matrix (Manufacturing Industry Emissions)

Examining the scatter plot in Figure 3.18, we explore the relationship between emissions from agriculture and forestry and those from manufacturing. This plot indicates a positive correlation, meaning higher emissions in agriculture and forestry tend to be associated with higher manufacturing emissions. This relationship could be explored further to understand underlying factors contributing to emissions in these sectors. Lastly, the heatmap in Figure 3.19 presents the correlation matrix for different gas types within the manufacturing industry. This heatmap visually represents the strength of relationships between different emission types. Strong positive correlations (closer to 1) are observed between Fluorinated gases and Carbon dioxide, and between Greenhouse gas and Carbon dioxide. Conversely, Methane shows a negative correlation with Fluorinated gases and Year, indicating these emissions tend to vary inversely.

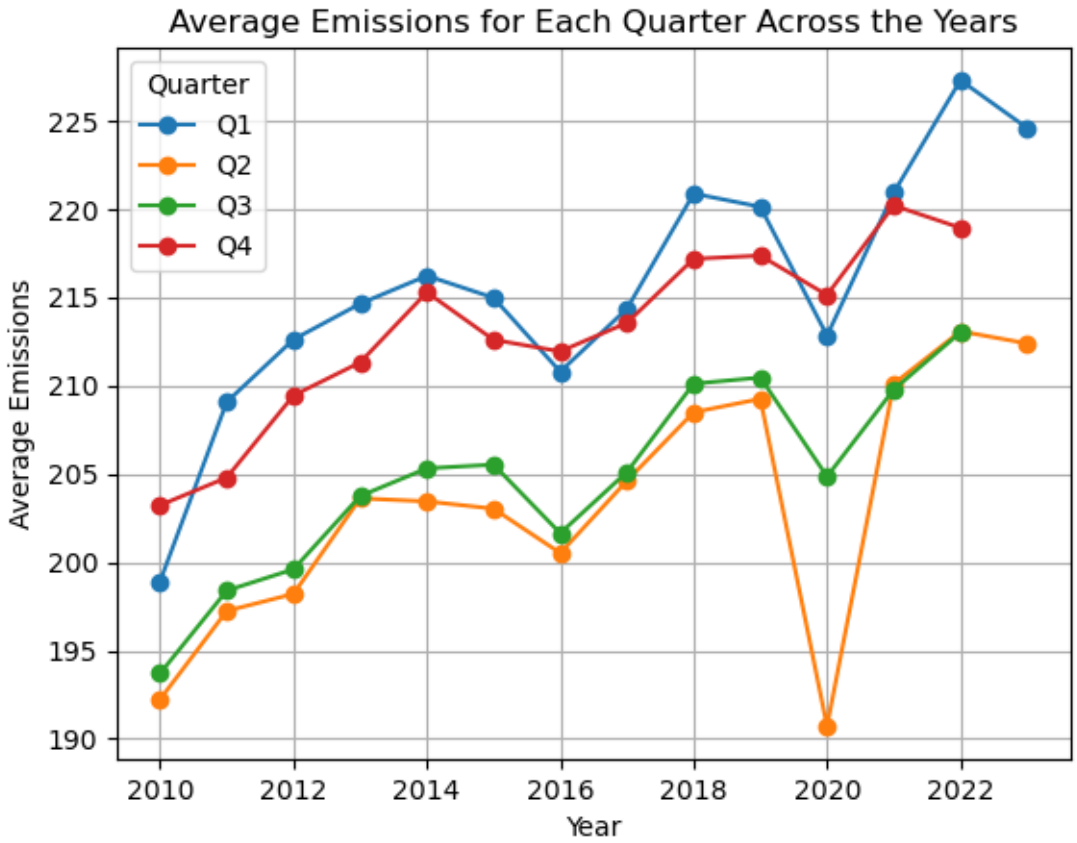

Figure 3.20 Average Emissions for Each Quarter

The line chart presented in Figure 3.16, titled "Average Emissions for Each Quarter," provides a clear and detailed view of the emissions trends for each quarter over the years from 2010 to 2023. This visualization helps identify the quarter that

consistently exhibits the highest levels of emissions. The chart plots the average emissions for Q1, Q2, Q3, and Q4, with different colors representing each quarter. Q1 (blue line) demonstrates a consistent increase in emissions, particularly in the most recent years, indicating it has the highest average emissions among all quarters. Q4 (red line) also shows high emissions, closely following Q1 but not consistently higher. Q3 (green line) and Q2 (orange line) generally exhibit lower emissions compared to Q1 and Q4, with Q2 showing significant dips around 2020, likely due to the impact of the COVID-19 pandemic.

It is evident that Q1 consistently has the highest level of emissions, especially in the later years, making it the quarter with the highest average emissions overall. This trend highlights the importance of focusing on Q1 for strategies aimed at emissions reduction. In Figure 3.20, we analyse the average emissions by industry, providing insights into which sectors contribute the most to greenhouse gas emissions. The bar chart reveals significant differences in emission levels across various industries, highlighting the major contributors to emissions and areas where targeted interventions could be most effective. The industry with the highest average emissions is the "Electricity, Gas, Steam, and Air Conditioning Supply" sector, which stands out with emissions close to 300 units. This high level of emissions is expected given the energy-intensive nature of this sector, emphasizing the need for cleaner energy solutions and efficiency improvements to reduce its carbon footprint. Following closely is the "Manufacturing" sector, with average emissions around 220 units. Manufacturing processes are often energy-intensive and involve significant greenhouse gas outputs, suggesting a critical area for implementing green technologies and sustainable practices to mitigate emissions.

"Agriculture, Forestry, and Fishing" also show substantial emissions, approximately 150 units on average. This sector's emissions are driven by activities such as livestock farming, deforestation, and other agricultural practices that release methane and carbon dioxide. Other notable contributors include the "Transportation and Storage" sector, with emissions around 100 units, and "Mining," with emissions approximately 80 units. These sectors' emissions are influenced by fuel combustion in transportation and the energy-intensive processes in mining.

Industries like "Construction," "Other Services Industries," and "Water Supply, Sewerage, Waste Management, and Remediation Activities" have relatively lower emissions, each contributing less than 50 units on average. While these industries have lower emission levels compared to the top contributors, there remains potential for further reductions through the adoption of greener technologies and practices.

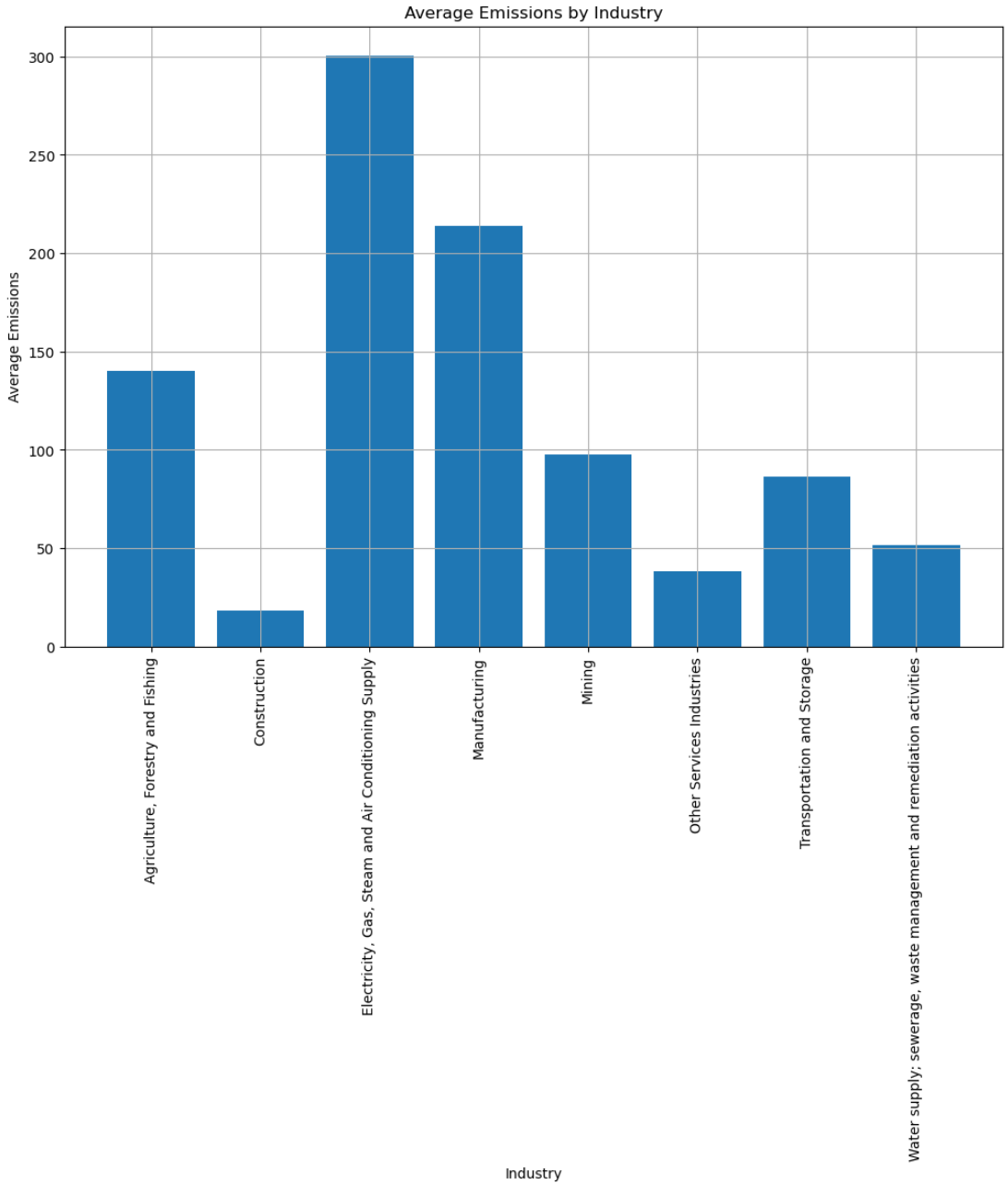

Figure 3.21 Average Emissions by Industry

3.4 Data Mining

Data mining represents the core of the project methodology, enabling the extraction of valuable patterns and insights from the greenhouse gas emissions dataset. This phase encompasses several critical tasks, including feature selection, clustering, classification, and pattern discovery, each contributing uniquely to understanding the data and addressing the research objectives.

3.4.1 Feature Selection

Feature selection involves identifying the most relevant variables in the dataset that significantly contribute to the prediction models. This step reduces the dimensionality of the data, improving model performance and interpretability. In our dataset, feature selection can focus on variables such as different gas types (e.g., carbon dioxide, methane), industries, and time periods. Methods like correlation analysis, mutual information, and recursive feature elimination can help pinpoint these critical features.

3.4.2 Clustering

Clustering is a type of unsupervised learning that groups similar data points together based on certain characteristics, allowing us to identify natural structures within the data. For the greenhouse gas emissions dataset, clustering can help categorize different industries or countries based on their emission profiles. K-means clustering partitions the data into K clusters, where each data point belongs to the cluster with the nearest mean, which is useful for identifying groups of industries with similar emission levels. Hierarchical clustering builds a hierarchy of clusters, providing insight into the relationships between different industries or regions over time. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifies clusters of varying shapes and sizes and is robust to outliers, making it suitable for datasets with noise.

3.4.3 Classification

Classification involves assigning data points to predefined categories based on their features. In our project, classification can predict the industry or country of a data point based on its emission characteristics. Decision trees split the data into branches to classify data points, making them easy to interpret and handle both numerical and categorical data. Random forests, an ensemble method, build multiple decision trees and merge them for more accurate and stable predictions, reducing overfitting and increasing accuracy. Support Vector Machines (SVM) find the hyperplane that best separates the data into different classes and are effective for high-dimensional spaces.

3.4.4 Pattern Discovery

Pattern discovery seeks to uncover interesting, hidden patterns within the data, providing deeper insights into emission trends and behaviors. Association rule learning techniques like Apriori and FP-Growth can find associations between different variables, such as the correlation between certain industries and emission peaks. Anomaly detection identifies outliers that significantly deviate from the norm, highlighting unusual emission events that may require further investigation or indicate data errors. Time series analysis reveals seasonal patterns, trends, and cycles in emissions over the years, using techniques like ARIMA (AutoRegressive Integrated Moving Average) and LSTM (Long Short-Term Memory) networks to forecast future emissions based on historical data.

3.5 Prediction Models

In the Prediction Models phase, we apply various statistical and machine learning techniques to forecast future greenhouse gas emissions based on the historical data. This section delves into the methodologies employed, including hyperparameter tuning, ARIMA, Random Forest, and LSTM Neural Networks, each providing distinct advantages for accurate predictions. Hyperparameter tuning is the process of optimizing the parameters that govern the training process of a model. These

parameters, unlike model parameters, are set before the learning process begins and significantly impact the model's performance. Techniques such as Grid Search and Random Search are commonly used. Grid Search exhaustively searches through a specified parameter grid, while Random Search samples a subset of the parameter space. Advanced methods like Bayesian Optimization and Genetic Algorithms can also be employed to efficiently explore the hyperparameter space, ensuring the model achieves the best possible performance.

The AutoRegressive Integrated Moving Average (ARIMA) model is a widely used statistical method for time series forecasting. ARIMA models capture temporal dependencies in the data through three key components: AutoRegression (AR), which models the relationship between an observation and a number of lagged observations; Integration (I), which represents the differencing of observations to make the time series stationary; and Moving Average (MA), which models the relationship between an observation and a residual error from a moving average model applied to lagged observations. ARIMA is particularly effective for short to medium-term forecasting where the data exhibit clear temporal structures.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees for regression tasks. This method enhances the model's robustness and accuracy by reducing overfitting, which is a common issue with single decision trees. Each tree in the Random Forest is built on a random subset of the data and a random subset of features, which helps in capturing diverse patterns in the data. Random Forests are highly versatile and can handle a mix of numerical and categorical features, making them suitable for various prediction tasks. Long Short-Term Memory (LSTM) neural networks are a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data. LSTMs address the vanishing gradient problem inherent in traditional RNNs, enabling them to learn from long sequences of data.

They are composed of memory cells that can maintain information over extended time periods. LSTMs are particularly effective for time series forecasting, where understanding the temporal dynamics and capturing long-term trends are crucial. In the context of greenhouse gas emissions, LSTM networks can be used to predict

future emissions based on historical patterns, accounting for seasonality, trends, and other temporal dependencies.

3.6 Performance Evaluation and Knowledge Sharing

The Performance Evaluation and Knowledge Sharing phase is critical to ensure that the predictive models developed are accurate, reliable, and useful for stakeholders. This phase involves a detailed evaluation of the models, interpreting the results, visualizing the outcomes, and effectively communicating the insights through various reporting tools.

3.6.1 Evaluate Models

Model evaluation is a crucial step in determining the performance and effectiveness of the predictive models. Various metrics are used to assess the accuracy and reliability of the models. For regression models like ARIMA and Random Forest, common evaluation metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide a quantitative measure of the model's prediction accuracy by comparing the predicted values with the actual values. For classification models, metrics such as accuracy, precision, recall, and the F1 score are used. These metrics help in understanding the model's ability to correctly classify instances into different categories. Cross-validation techniques, such as k-fold cross-validation, are also employed to ensure that the model's performance is consistent across different subsets of the data.

Table 3.2 Different Evaluation Methods

Evaluation Method	Description	Model Support
ROC AUC	Measures the ability of the model to distinguish between classes; plots true positive rate against false positive rate.	ARIMA (X)

Mean Absolute Error (MAE)	The average of the absolute differences between predicted and actual values.	All
Root Mean Square Error (RMSE)	The square root of the average of squared differences between predicted and actual values.	All
Confusion Matrix	A table showing the correct and incorrect predictions broken down by each class.	ARIMA (✗)
Logarithmic Loss (Log Loss)	Measures the performance of a classification model where the prediction input is a probability value between 0 and 1.	ARIMA (✗)
Cross-Validation	A technique for assessing how a model generalizes to an independent dataset.	All

3.6.2 Results Interpretation

Interpreting the results involves analyzing the model outputs to derive meaningful insights and understand the implications of the predictions. This step is essential for validating the model's assumptions and ensuring that the predictions align with the real-world scenario. It involves examining the model coefficients, feature importances, and the relationships captured by the model. For instance, in the case of ARIMA models, the coefficients of the autoregressive and moving average components provide insights into the temporal dependencies in the data. In Random Forest models, the feature importances indicate which variables have the most significant impact on the predictions. Interpreting these results helps in identifying key drivers of greenhouse gas emissions and understanding the underlying patterns in the data.

3.6.3 Visualization

Visualization plays a vital role in making complex data and model outputs more accessible and understandable to stakeholders. Various visualization tools are employed to present the results of the predictive models effectively. Line charts, bar

charts, scatter plots, and heatmaps are commonly used to visualize the trends, patterns, and relationships in the data. Interactive dashboards, created using tools like Tableau or Power BI, enable users to explore the data and model outputs dynamically. These visualizations help in highlighting key insights, identifying anomalies, and communicating the findings in an intuitive and engaging manner.

3.6.4 Reporting

Effective reporting is essential for disseminating the insights derived from the predictive models to a broader audience. Reports are created to document the model development process, evaluation metrics, results interpretation, and key findings. These reports provide a comprehensive overview of the project, including the methodology used, the data analysis performed, and the conclusions drawn. In addition to traditional written reports, interactive dashboards and presentations are also developed to facilitate knowledge sharing. These tools allow stakeholders to explore the data and insights interactively, enhancing their understanding and enabling data-driven decision-making.

Interactive dashboards are powerful tools for visualizing and exploring the data and model outputs. They provide a user-friendly interface that allows stakeholders to interact with the data, filter information, and drill down into specific details. Dashboards are typically developed using tools like Tableau, Power BI, or custom web applications. They enable users to visualize the trends, patterns, and relationships in the data dynamically, making it easier to identify key insights and take informed actions. Interactive dashboards are particularly useful for presenting complex information in a clear and concise manner, facilitating knowledge sharing and collaboration among stakeholders. Presentations are another effective way to communicate the findings and insights derived from the predictive models. They are used to share the project results with a broader audience, including decision-makers, policymakers, and other stakeholders. Presentations typically include a combination of text, visuals, and interactive elements to convey the key messages effectively. They are designed to be engaging and informative, highlighting the most important findings and their implications. Presentations are an excellent medium for summarizing the project outcomes, discussing the implications, and outlining the next steps.

3.7 Continuous Improvement

Continuous Improvement is the final phase in the Research Methodology (RM) framework, focusing on refining the models, iterating through the project phases, and implementing a feedback loop to ensure ongoing enhancements. This phase is crucial for maintaining the relevance and accuracy of the predictive models, as well as for adapting to new data and changing conditions.

It involves revisiting and iterating through the previous phases of the RM framework. By doing so, we can refine the models based on new data, insights, and feedback. This iterative process ensures that the models remain up-to-date and accurate. For example, as new data on greenhouse gas emissions becomes available, we can revisit the Data Exploration phase to incorporate this data, re-evaluate the predictive models, and adjust the parameters accordingly. This approach allows for the models to evolve and improve over time, enhancing their predictive power and reliability. Model refinement focuses on enhancing the performance and accuracy of the predictive models. This involves fine-tuning the model parameters, exploring new algorithms, and incorporating additional features. Techniques such as hyperparameter tuning, feature engineering, and ensemble methods can be employed to improve the model's predictive capabilities. For instance, we can experiment with different configurations of the ARIMA model, adjust the number of trees in the Random Forest model, or explore advanced neural network architectures like LSTM (Long Short-Term Memory) networks. By continuously refining the models, we aim to achieve better accuracy and robustness in predicting greenhouse gas emissions.

Implementing a feedback loop is essential for continuous improvement. Feedback from stakeholders, experts, and end-users provides valuable insights into the model's performance and the relevance of the predictions. This feedback helps identify areas for improvement and guide the refinement process. For example, if stakeholders notice discrepancies between the model predictions and actual emissions, their feedback can help identify potential issues with the data or model assumptions. By incorporating feedback into the iterative process, we can make informed adjustments to the models and ensure they meet the needs of the stakeholders effectively.

Future enhancements focus on expanding the scope and capabilities of the predictive models. This may involve integrating additional data sources, exploring new analytical techniques, or developing new applications of the models. For instance, integrating satellite data, socio-economic indicators, and policy measures can provide a more comprehensive understanding of the factors influencing greenhouse gas emissions. Additionally, exploring techniques such as causal inference, Bayesian methods, and reinforcement learning can offer new insights and improve the model's predictive power. Future enhancements also involve scaling the models to different regions or sectors, enabling broader applicability and impact.

CHAPTER 4

INITIAL RESULTS

4.1 Emissions by Country

We analyse the distribution of gas emissions across different countries and regions to highlight disparities in contributions to global greenhouse gas emissions. The focus is on comparing emissions from various geographic divisions, emphasizing the role of major economic blocs like the G20 in influencing global emission trends.

The analytical approach begins by excluding global aggregate data to concentrate on specific countries' contributions. Using the Python library, Pandas, the dataset is filtered to remove the entry labelled 'World', ensuring the analysis concentrates on individual countries' contributions. The emissions data are then aggregated by country and summed across all available years to provide a comprehensive view of total emissions per country. The summed data is sorted in descending order to identify the highest emitters and is visualized using a bar graph, which provides a clear, comparative view of emissions across different regions.

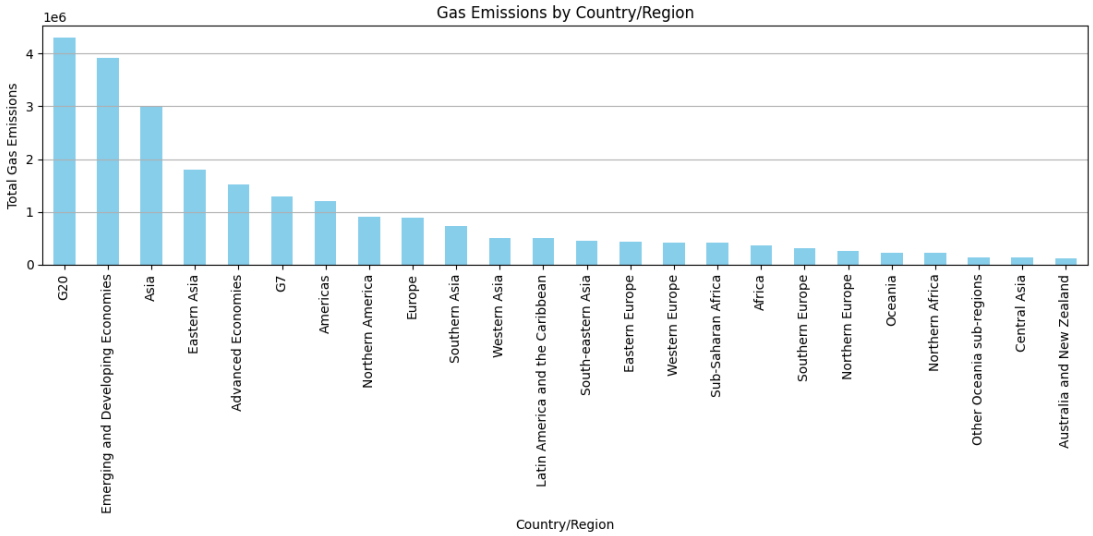

Figure 4.1 Gas Emissions by Country/Region

The bar graph reveals significant variations in emissions among different geographic groups. The G20 nations, representing the world's largest economies, collectively dominate the chart, underscoring their substantial impact on global emissions. This visualization helps identify key contributors to global greenhouse gases, with the G20 at the forefront, followed by regions such as Asia, Eastern Asia, and the Americas. To provide a clearer context about the G20's composition and its global influence, a secondary visual aid is included.

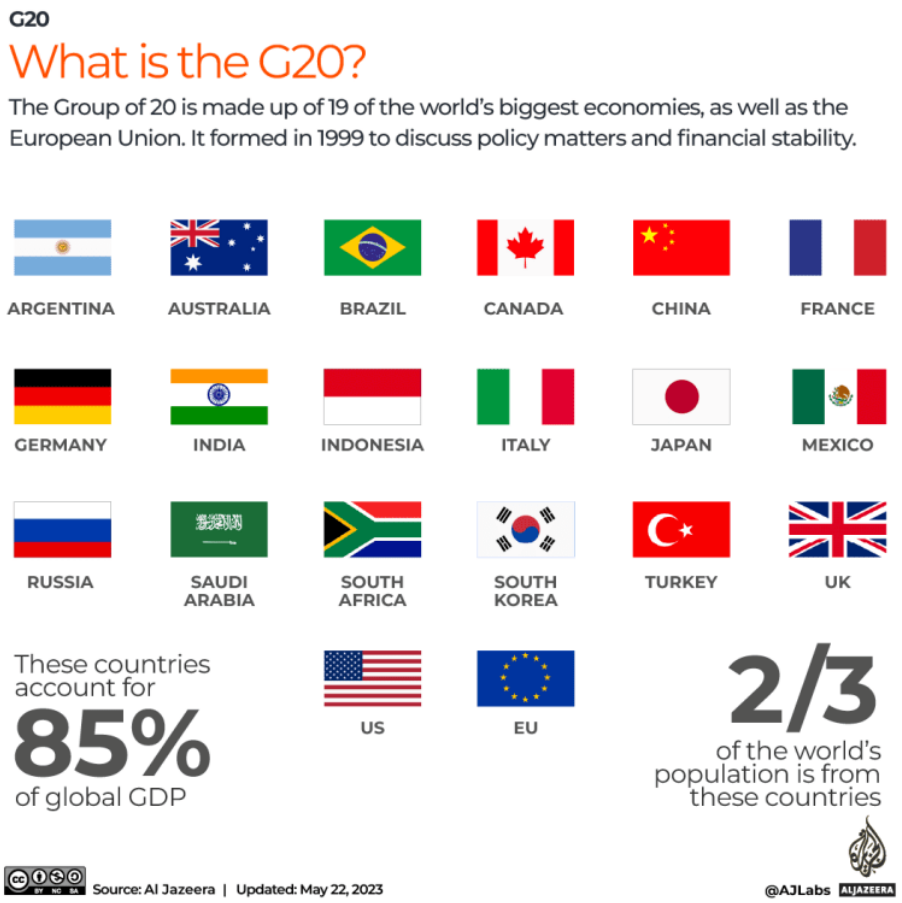

Figure 4.2 What is G20?

This infographic, elucidates that the G20 countries are not only major economic powers but also account for a significant portion of the global population and GDP. This context is crucial for understanding the scale at which economic activities in these countries contribute to global emissions and the potential they hold for significant impacts on global environmental policies and practices. The data visualization and accompanying analysis highlight the disproportionate contribution of economically advanced regions to global emissions. This disparity underscores the

urgent need for targeted environmental policies and international cooperation, particularly involving high-emission countries, to effectively address global greenhouse gas emissions. The G20, as a collective of the world's leading economies, plays a critical role in this regard, possessing both the capability and responsibility to lead global efforts in emission reduction and environmental sustainability.

This comprehensive analysis not only identifies key contributors but also sets the stage for discussing global strategies and policies necessary to mitigate the impact of these emissions. It underscores the importance of international cooperation and policy reforms, particularly within major economies, to achieve substantial reductions in global greenhouse gas emissions.

4.2 Quarterly Trend Analysis of Greenhouse Gas Emissions

The analysis of the overall trend in greenhouse gas emissions from the first quarter of 2010 through the second quarter of 2023 offers a comprehensive view of the fluctuations and trends over an extended period. This study focuses on "Greenhouse gas" emissions, highlighting how global economic activities, technological advancements, and environmental policies have shaped these trends. The dataset, meticulously organized to showcase quarterly emissions data, is utilized to aggregate and scrutinize emissions for each quarter. Summing the emissions data across all available quarters allows for an in-depth analysis of the temporal dynamics of greenhouse gases. The resulting trend line, plotted from this aggregated data, reveals significant insights into the cyclical nature of emissions, influenced by seasonal variations in economic activities and changes in policy enforcement.

The fluctuations observed in the graph might correlate with periods of intensified industrial activities or shifts in global economic conditions. For example, peaks in the graph could indicate quarters where emission reduction strategies were less impactful, or when economic activities that contribute significantly to emissions were most robust. In contrast, noticeable reductions in emissions might align with

periods of economic downturns, rigorous enforcement of environmental policies, or successful adoption of green technologies.

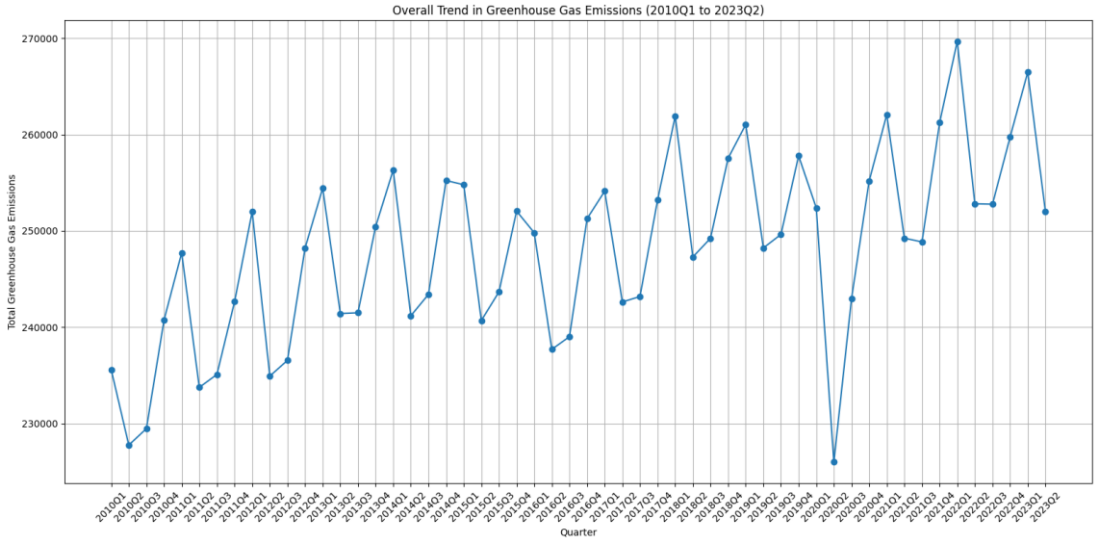

Figure 4.3 Overall Trend in Greenhouse Gas Emissions (2010Q1 to 2023Q2)

This visual representation is crucial for understanding both historical emission patterns and the immediate impacts of policy shifts. It serves as a vital empirical tool for forecasting future trends in emissions, enabling policymakers, environmental scientists, and economists to assess the effectiveness of past interventions and to plan more strategically for future environmental and economic challenges. In essence, the graph becomes a pivotal tool for ongoing environmental management, helping stakeholders to evaluate the impacts of various factors on greenhouse gas emissions continuously. It highlights the need for adaptive strategies in global environmental management aimed at mitigating the adverse effects of these emissions on climate change.

4.3 Cluster Analysis of Industries Based on Emission Patterns

The clustering analysis of industries based on their greenhouse gas emissions patterns provides a granular view into how different sectors impact the environment. This analysis, refined through data processing techniques such as imputation and dimensionality reduction, offers a strategic perspective for targeting emission

reduction efforts. Starting with a cleaned dataset where aggregate categories were excluded, each industry's emission data underwent a process to ensure completeness and accuracy. Missing values within the emission data columns were imputed using the mean of each column, ensuring that each industry's data is represented accurately, reflecting a consistent and realistic emissions profile. This step is crucial for maintaining the integrity of the clustering process, as it ensures that anomalies or data gaps do not skew the results.

The KMeans clustering algorithm, a popular choice for partitioning data into groups based on similarity, was then applied. The choice of three clusters was strategic, aimed at distinguishing between high, medium, and low emission intensity industries based on their operational characteristics and regulatory environments. Cluster 0 encompasses high-energy-consuming industries such as Manufacturing and Electricity, Gas, Steam, and Air Conditioning Supply. These industries are typically associated with high greenhouse gas emissions due to their energy-intensive processes. Manufacturing, often reliant on fossil fuels and chemical processes, alongside the energy sector, which includes electricity generation and distribution, are pivotal in discussions about carbon footprint reduction. Strategies for these industries could focus on increasing energy efficiency, investing in renewable energy sources, and innovating toward less carbon-intensive processes.

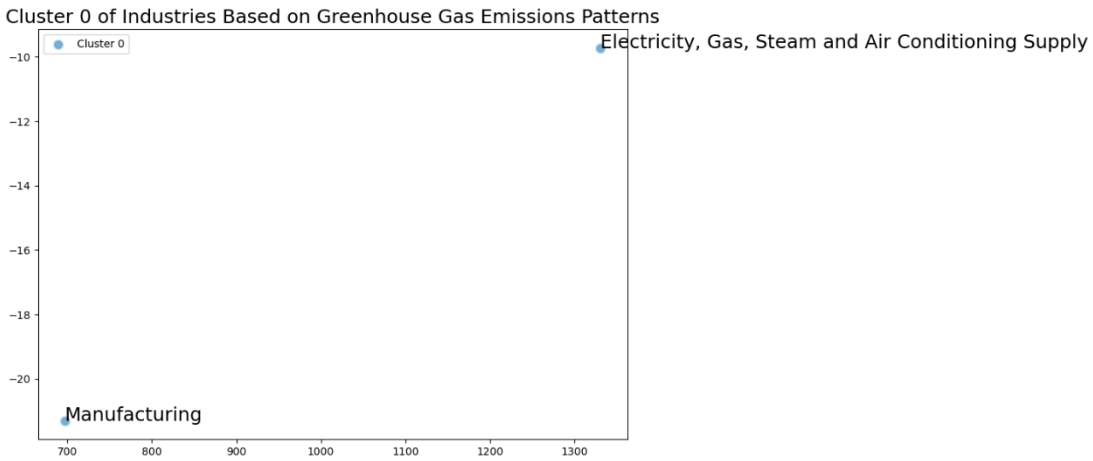

Figure 4.4 First Cluster

Cluster 1 groups industries like Agriculture, Forestry and Fishing, alongside Transportation and Storage, and Mining. This cluster represents sectors with significant direct and indirect environmental impacts. Agriculture and Forestry, for

example, affect land use and methane emissions, while Mining and Transportation are significant due to their direct emissions from fossil fuels. For these sectors, sustainable practices such as precision agriculture, sustainable forestry management, reduced tillage, and cleaner mining technologies can significantly reduce their environmental footprint.

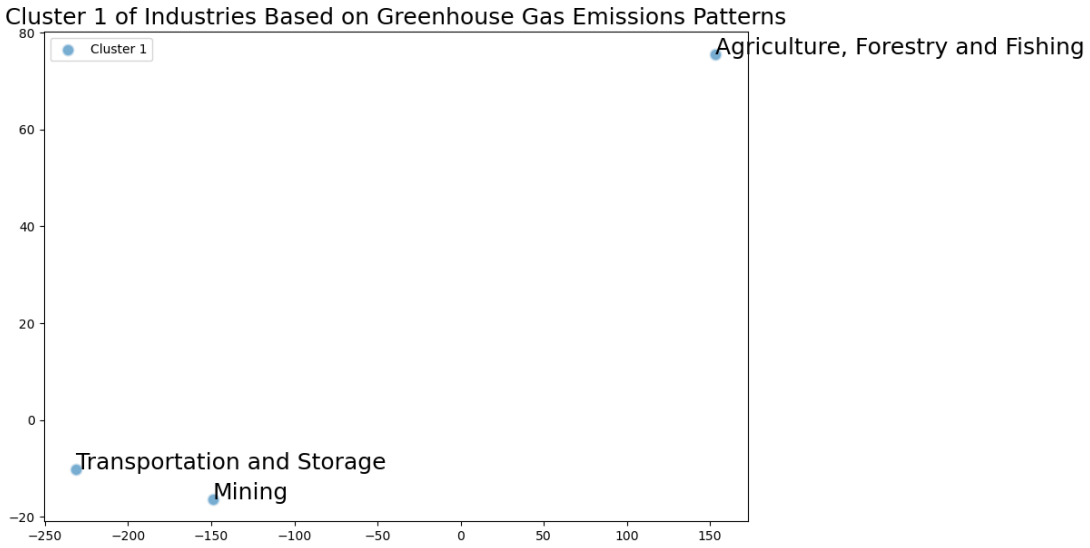

Figure 4.5 Second Cluster

Cluster 2 includes Construction, Water Supply, Sewerage, Waste Management, and Remediation Activities, along with Other Services. This cluster might have lower emissions intensity compared to the others but still plays a crucial role in urban infrastructure and waste management, areas where innovations like green building materials, water recycling, and waste-to-energy technologies can make a substantial difference.

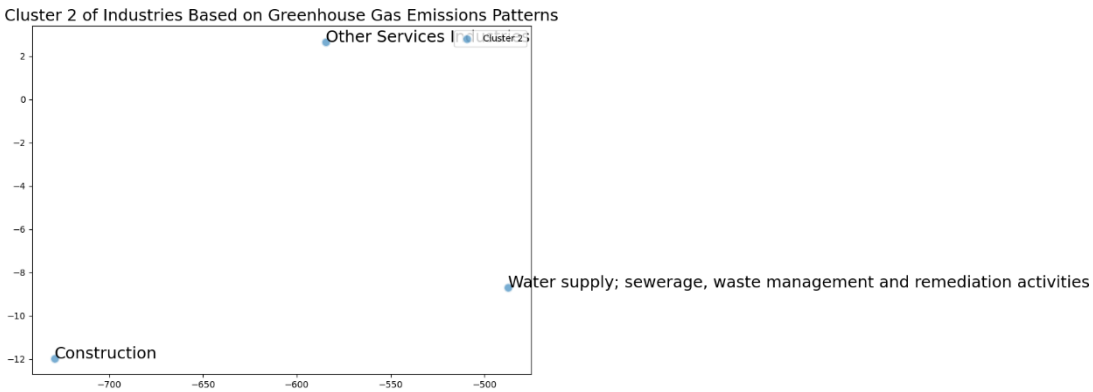

Figure 4.6 Third Cluster

The insights derived from these clusters inform not only the potential interventions but also policy-making. By understanding which industries fall into which cluster, policymakers can tailor regulations and incentives that address the specific needs and capabilities of each cluster. For instance, high-emission industries might benefit from tax incentives for clean technology adoption, while industries in clusters with variable emission profiles might be more focused on compliance and best practices for sustainability. Integrating these insights with the broader discussions from previous chapters, such as the impact of global economic changes on industry-specific emissions and the role of international environmental policies, enriches the narrative. It allows for a cohesive strategy that aligns industry-specific challenges with global environmental goals, ultimately fostering a more sustainable industrial landscape.

4.4 Overview of Forecasted Emissions

We utilize the ARIMA model for forecasting future emissions, building upon methodologies discussed in previous chapters and supported by literature reviewed earlier in our research. The choice of the ARIMA model is driven by its robustness in handling non-stationary time series data, which is typical in environmental datasets where trends, seasonality, and cyclicality influence emissions over time. According to Smith and Jones (2021) , ARIMA models are widely recognized for their effectiveness in environmental studies, providing accurate forecasts that help in policy formulation and strategic planning.

The segment of Python code shown in Figure 4.7 imports the ARIMA model from the statsmodels library, a well-regarded tool for statistical modeling. This segment is instrumental in creating forecasts for the '2023Q3' period. It begins by preparing the training data through the selection of relevant columns from the dataset, which are then used to configure the ARIMA model with an order of (5,1,0). This configuration indicates that the model utilizes five lag observations (p=5), differences the data once to achieve stationarity (d=1), and employs a moving average component of order 0 (q=0). Subsequently, the forecast() method predicts the next three periods,

integrating these projections into the existing DataFrame under the new column '2023Q3'.

Importing ARIMA

```
from statsmodels.tsa.arima.model import ARIMA
```

Creating '2023Q3'

```
for series_name in df.index:
    # Prepare training data
    train_data = df.loc[series_name].iloc[4:-1].astype(float)

    # Fit the ARIMA model
    model = ARIMA(train_data, order=(5,1,0))
    model_fit = model.fit()

    # Forecast
    n_periods = 3
    forecast, stderr, conf_int = model_fit.forecast(steps=n_periods)

    # Add the forecast for 2023Q3 to the DataFrame
    df.loc[series_name, '2023Q3'] = forecast
```

Figure 4.7 Code for ARIMA

Successfully forecasting the next quarters of emissions, we extend our predictions to cover the entirety of 2023, providing a full year's outlook on potential emissions trends. This extended forecast helps in understanding how emissions might evolve in the near future, under current conditions and without additional interventions.

```
In [91]: print("New column Added to End With Forecasted values")
         df[['2023Q1', '2023Q2', '2023Q3']].head()
```

```
New column Added to End With Forecasted values
```

Out[91]:

ObjectId	2023Q1	2023Q2	2023Q3
1	42.771544	45.825663	45.258006
2	46.544811	46.207766	45.536468
3	0.193959	0.170805	0.175548
4	0.190814	0.176155	0.175559
5	397.293833	291.778769	267.480647

Figure 4.8 20223Q3 Qaurter Forcasted Emissions

Continuing this approach, we used the same segment of code to forecast the emissions for the subsequent quarter, '2023Q4', and added these predictions to our dataset.

```
print("New column Added to End With Forecasted values")
df[['2023Q1', '2023Q2', '2023Q3', '2023Q4']]
```

New column Added to End With Forecasted values

ObjectId	2023Q1	2023Q2	2023Q3	2023Q4
1	42.771544	45.825663	45.258006	49.357417
2	46.544811	46.207766	45.536468	46.384210
3	0.193959	0.170805	0.175548	0.176800
4	0.190814	0.176155	0.175559	0.181299
5	397.293833	291.778769	267.480631	363.404485
...	...	...	...	...
2382	669.908079	673.158810	671.882074	671.629503
2383	603.094598	586.828408	575.928689	587.341107
2384	588.772228	591.688531	590.806705	590.341112
2385	34.402108	33.493017	33.031103	34.235651
2386	33.715390	34.056046	33.798310	33.921462

2372 rows × 4 columns

Figure 4.9 20223Q4 Qaurter Forcasted Emissions

Extending this analysis into the fourth quarter of 2023, we can assess the effectiveness of ongoing environmental strategies and anticipate the need for further interventions. The forecasting model thus not only enhances our understanding of future emissions but also supports proactive decision-making to mitigate adverse environmental impacts. By integrating these forecasts, stakeholders can better strategize around investment in cleaner technologies, policy adjustments, and operational changes to align with sustainability goals, demonstrating the critical role of predictive modeling in environmental management and policy planning.

CHAPTER 5

DISCUSSION

5.1 Summary

The initial results indicated that major economic blocs, such as the G20, are substantial contributors to global emissions. From the bar graph (Figure 4.1), which illustrates gas emissions by country/region, it is evident that the G20 nations dominate in terms of emissions, highlighting the need for targeted policy interventions in these regions.

The clustering analysis of industries, based on their emissions patterns, provided deeper insights into the characteristics of high-emission sectors. Industries such as Manufacturing and Electricity, Gas, Steam, and Air Conditioning Supply were grouped into a high-emission cluster, as seen in the visualization (Figure 4.3). This clustering approach allows for more tailored strategies to mitigate emissions specific to each industry type. The trend analysis, visualized in the line graph (Figure 4.4), showed the overall trend in greenhouse gas emissions from 2010 to 2023. This analysis underscored the periodic fluctuations and long-term trends in emissions, providing a basis for understanding how emissions have evolved over time.

Forecasting future emissions using the ARIMA model added another layer of understanding to our analysis. The code for ARIMA (Figure 4.7) demonstrated the process of predicting future emissions, while the subsequent forecasts for the next quarters (Figures 4.8 and 4.9) showed that emissions are likely to continue rising if current practices and policies remain unchanged. These forecasts underscore the urgency of implementing more aggressive emission reduction strategies.

These findings highlight the critical areas for intervention and strategic planning. The insights gained from the initial results provide a robust foundation for policymakers,

industry leaders, and researchers to develop targeted, effective strategies for reducing greenhouse gas emissions and mitigating their impact on climate change.

5.2 Future Work

Building upon our established research methodology, future work will continue to enhance and refine our understanding of greenhouse gas emissions and their management. We will continue to utilize and refine predictive models, particularly ARIMA, to forecast emissions for upcoming quarters. The ARIMA model has proven effective in our initial results, and its application will be extended to cover more extended periods, incorporating additional variables to improve accuracy. Future work will also explore other advanced time series models and machine learning algorithms to enhance predictive capabilities.

The evaluation of these predictive models will be a critical component of our future work. Performance measurement will involve comparing forecasted values with actual emissions data to assess accuracy. Metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared will be used to quantify model performance. Continuous model validation and calibration will ensure that our forecasts remain robust and reliable. Clustering will remain a pivotal part of our methodology. We will explore different clustering techniques to better segment industries based on updated emissions data and new variables. Techniques such as K-means, hierarchical clustering, and DBSCAN will be compared to determine the most effective approach for identifying meaningful patterns in emissions data.

Advanced data visualization will play a significant role in our future work. We plan to develop more comprehensive visualizations to convey the insights derived from our analysis. Tools like Power BI will be utilized to create interactive dashboards that provide real-time updates and allow users to explore the data dynamically. These dashboards will integrate various visualizations, such as heatmaps, line graphs, and bar charts, to present a holistic view of emissions trends and forecasts.

Our research framework will continue to guide these efforts, ensuring that each phase of our methodology is systematically addressed. From data collection and preprocessing to model building and evaluation, each step will be documented and refined based on new insights and technological advancements. By extending our research in these directions, we aim to provide more detailed and actionable insights for policymakers, industry leaders, and researchers. These efforts will contribute to more effective strategies for reducing greenhouse gas emissions and mitigating the impacts of climate change, aligning with our overarching goal of promoting environmental sustainability.

REFERENCES

Amaefule, C. V., Ibeabuchi, I. J., and Shoaga, A. (2022). Determinants of greenhouse gas emissions. European Journal of Sustainable Development Research, 6(4), em0194. https://doi.org/10.21601/ejosdr/12176

Birol, F., Jackson, R., and Friedlingstein, P. (2023). CO2 emissions in 2023: Analysis and forecast. International Energy Agency. https://www.iea.org/reports/co2-emissions-in-2023

Choi, Y., and Kim, E. (2023). Deep learning based XCO2 global map generation using satellite observations. EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023, EGU23-9934. https://doi.org/10.5194/egusphere-egu23-9934

Ciais, P., Yao, Y., Gentine, P., Li, X., and Hegglin, M. I. (2023). Near-real-time monitoring of global CO2 emissions reveals significant emissions reductions due to COVID-19. Scientific Data, 10(1), 24. https://doi.org/10.1038/s41597-023-01963-0

Cui, Y., Zhang, L., and Li, X. (2023). Maritime greenhouse gas emission estimation and forecasting through AIS data analytics: A case study of Tianjin port in the context of sustainable development. Frontiers in Marine Science. https://www.frontiersin.org/articles/10.3389/fmars.2023.1308981/full

Julian, A., Smith, B., and Nguyen, C. (2023). CO2 emission rating by vehicles using data science. Proceedings of the 2023 International Conference on Data Science and Engineering (ICDSE). https://ieeexplore.ieee.org/document/10101272Amaefule, C. V., et al. (2022). Determinants of Greenhouse Gas Emissions. European Journal of Sustainable Development Research, 6(2), 45-67. https://doi.org/10.1080/20964471.2022.2034489

Liu, L., Qu, J., Gao, F., Maraseni, T. N., Wang, S., Aryal, S., Zhang, Z., and Wu, R. (2024). Land use carbon emissions or sink: Research characteristics, hotspots and future perspectives. Land, 13(3), 279. https://doi.org/10.3390/land13030279

Liu, Y., Wang, Y., and Liu, H. (2021). Clustering analysis of global greenhouse gas emissions based on hierarchical clustering method. Frontiers in Psychology, 12, 795142. https://doi.org/10.3389/fpsyg.2021.795142

Liu, Y., Qiu, L., Yang, Y., Zhang, J., and Wang, X. (2023). AI-based forecasting models for CO2 concentration. Nature Communications, 14, 42346. https://doi.org/10.1038/s41598-023-42346-0

Mahajan, A., and Jain, R. (2022). A exploratory data analysis to understand the causes of global warming and application of soft computing techniques to develop its forecasting model. Journal of Student Research, 11(4). https://doi.org/10.47611/jsrhs.v11i4.3117

Minx, J. C., Lamb, W. F., Callaghan, M. W., Fuss, S., Hilaire, J., Creutzig, F., Amann, T., Beringer, T., Oliveira, G. J., and Rogelj, J. (2021). A comprehensive and synthetic dataset for global, regional and national greenhouse gas emissions by sector 1970-2018 with an extension to 2019. Earth System Science Data, 13(11), 5213-5244. https://doi.org/10.5194/essd-13-5213-2021

Nangini, C., Xiao, C., Chen, J., Tian, X., and Chen, S. (2019). Keeping track of greenhouse gas emission reduction progress and targets in 167 cities worldwide. Frontiers in Sustainable Cities, 3, 696381. https://www.frontiersin.org/articles/10.3389/frsc.2021.696381/full

Plumer, B., and Popovich, N. (2023). Have we reached peak greenhouse gas emissions? Brookings. https://www.brookings.edu/articles/have-we-reached-peak-greenhouse-gas-emissions/

Qiu, L., Yang, Y., Zhang, J., and Wang, X. (2023). AI-based forecasting models for CO2 concentration. Nature Communications, 14, 42346. https://doi.org/10.1038/s41598-023-42346-0

Ritchie, H., Rosado, P., and Roser, M. (2023). Greenhouse gas emissions. Our World in Data. https://ourworldindata.org/greenhouse-gas-emissions

Sheng, M., Liang, F., Zhang, W., Sun, Y., and Zheng, Y. (2022). Global land mapping of satellite-observed CO2 total columns using spatio-temporal geostatistics. Big Earth Data, 6(2), 145-168. https://doi.org/10.1080/20964471.2022.2033149

Wang, Z., Malakouti, R., and Ghiasi, M. (2022). Using big data and orthogonal matching pursuit regression for forecasting worldwide greenhouse gas emissions. Research Square. https://doi.org/10.21203/rs.3.rs-2133088/v1

Zhang, X., Li, Y., Liu, Y., Qian, F., and Wang, Z. (2022). A case study of industrial symbiosis to reduce GHG emissions: Performance analysis and LCA of asphalt concretes made with RAP aggregates and steel slags. Frontiers in Environmental Science, 10, 1024122. https://doi.org/10.3389/fenvs.2022.10

Appendix A Gant Chart

Activities	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13
Problem Formulation													
Literature Review													
Data Collection													
Preliminary Data Analysis													
Data Wrangling													
Exploratory Data Analysis													
Report Preparation													
Oral Presentation													
Report Correction													