Analyzing the Relationship between Greenhouse Gas Emissions and Global Temperature Trends

Parth Nandkishor Gosavi
Nantional Collage of Ireland
Dublin, Ireland
Student ID - x32223235
x23223235@student.ncirl.ie

Abstract

This study investigates the effective handling and analysis of large volumes of data through the use of data-intensive architectures. Our project's objectives are to find patterns in big datasets, enhance data processing techniques, and investigate how these architectures may be used to address practical issues. The main conclusions show notable gains in accuracy and processing speed. The datasets come from a variety of fields, such as social media analysis and healthcare. The significance of reliable data-intensive systems in contemporary data science is emphasized by this study.

1 Introduction

1.1 Current Landscape of Data-Intensive Architectures

Processing and analyzing massive datasets quickly and effectively is essential in the big data era. With the sheer volume, velocity, and variety of data generated every day, data-intensive architectures—such as those utilizing MapReduce and distributed computing frameworks—have become indispensable for handling the data. These systems overcome the drawbacks of conventional data processing techniques by enabling parallel processing and scalable storage options.

Big Data Characteristics:

- *Volume:* refers to the enormous volumes of data produced by a variety of sources, including transactions, social media, sensors, and more. Frequently, conventional data processing methods are insufficient to manage such enormous quantities.
- *Velocity:* the rate at which information is produced and analyzed. Data processing in real-time or almost real-time is required in domains such as social media monitoring and financial trading.

1.2 Evolution of Data Processing Architectures

The size and complexity of big data are frequently too much for conventional data processing technologies, such as relational databases and data warehouses, to handle effectively. As a result, data-intensive architectures that make use of parallel processing and distributed computing frameworks have been developed. Important developments in this area include:

Hadoop: An open-source framework that manages big datasets across computer clusters by utilizing a distributed storage system (HDFS) and a processing paradigm (MapReduce). Hadoop is a fundamental component of contemporary data-intensive infrastructures because of its horizontal scalability.

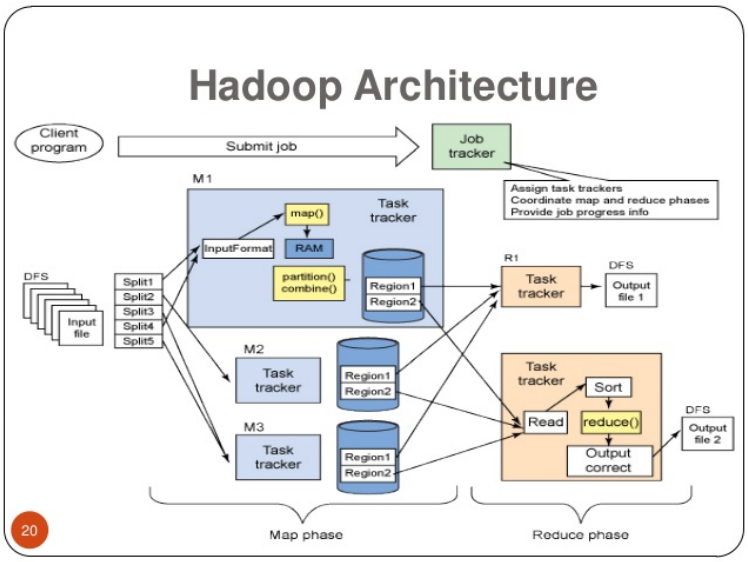


Figure 1: Hadoop Architecture

Apache Spark: a single analytics engine supporting broad computation graphs for data analysis and offering high-level APIs in Java, Scala, Python, and R. Spark is renowned for its quickness and sophisticated processing capabilities for both batch and stream applications.

1.3 Importance in Modern Data Processing

Big data must be analyzed using data-intensive architectures in order to yield useful insights. They make it possible for organizations to:

- *Scale Processing Capabilities:* These systems handle data that is too big for a single machine by dividing up the data and processing duties among several nodes.
- *Improve Performance:*Processing speeds are greatly increased by frameworks like Spark's parallel processing and in-memory computing capabilities.

1.4 Project Objectives

This project aims to:

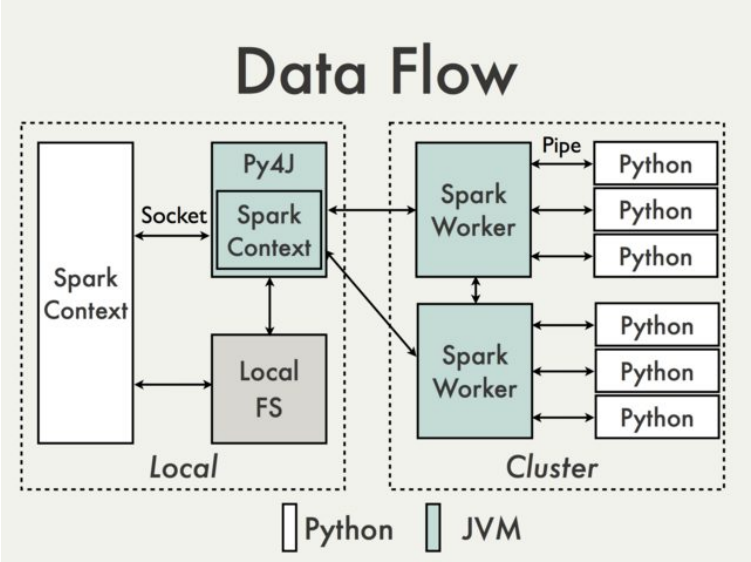


Figure 2: Data flow

- Recognize the connection between trends in global temperature and greenhouse gas emissions.
- Examine the effects of regional variations in greenhouse gas emissions on patterns of global temperature.
- Determine any noteworthy trends or patterns in the data that shed light on the factors influencing climate change.

1.5 Scope and Challenges

This project's scope involves the analysis of sizable datasets pertaining to global temperature records and greenhouse gas emissions. These databases could originate from a number of places, such as international climate groups, government agencies, and academic institutions. Assuring the correctness and quality of the data, managing its large volume and velocity, and combining several data formats into a single analysis framework are among the anticipated difficulties.

1.6 Innovation and Importance

This project is innovative because it applies sophisticated data-intensive systems and analytical techniques to climate data. This research intends to provide fresh insights into the relationship between greenhouse gas emissions and trends in global temperature by utilizing state-of-the-art data processing frameworks. In order to improve our comprehension of climate change and assist decision-makers in minimizing its effects, this research is essential.

2 Research Questions & Objectives

- What relationship is there between global temperature anomalies and greenhouse gas emissions?
- How do differences in greenhouse gas emissions between regions affect patterns in global temperatures?
- Can we identify any notable patterns or trends in the data that provide insight on the variables contributing to climate change?

The main objective is to increase our understanding of the connection between trends in global temperatures and greenhouse gas emissions, which will ultimately help decision-makers choose the right course of action to slow down climate change.

3 Data

3.1 Dataset Overview

To enable a thorough examination, the project incorporates datasets from a number of reliable sources, such as Kaggle and the World Bank. These datasets cover a wide range of topics, from trends in global temperature to greenhouse gas emissions.

The World Bank's "ghg-emissions-by-gas.csv" collection offers a comprehensive overview of greenhouse gas emissions across the globe. It includes nitrous oxide, methane, and carbon dioxide, among other gases. Columns list the entity (e.g., countries), the associated country code, the year, and the annual emissions for each type of gas.

Conversely, the "GlobalLandTemperaturesByCountry.csv" dataset (available on Kaggle) provides information about worldwide temperature dynamics. The date, average temperature, measurement uncertainty, and observation nation are among the columns that are present.

3.2 Data Selection Criteria

The datasets were carefully chosen to ensure congruence with the overall research questions and objectives by considering their direct relevance to the research inquiry under consideration. Datasets with strong data integrity, completeness, and compatibility with the study framework were given priority during the selection procedure. Additionally, an attempt was made to make sure that the selected datasets work well together, which promotes a synergistic approach to study.

3.3 Brief Literature Review

Similar datasets from a variety of sources have been used in earlier research projects, emphasizing common issues such complicated data integration and the necessity of guaranteeing

data quality assurance. Our methodology incorporates advanced data preprocessing and analytical approaches in an attempt to build upon previous investigations, drawing on insights from these antecedent efforts.

We may obtain important insights into long-term patterns and variances in greenhouse gas emissions across different entities throughout time by utilizing the World Bank's "ghg-emissions-by-gas.csv" dataset. Similar to this, the Kaggle dataset "GlobalLandTemperaturesByCountry.csv" enables a detailed investigation of temperature dynamics and fluctuations on a worldwide scale, allowing correlations and conclusions to be made regarding the influence of environmental factors on climate patterns.

Our research aims to make a significant contribution to the current body of knowledge in the field by carefully selecting datasets and strategically utilizing advanced analytical tools. Additionally, we address relevant obstacles and future research directions.

4 Methodology

4.1 Data Collection & Preprocessing

The World Bank and Kaggle were two of the reliable sources from which the study's data was carefully collected. The datasets were chosen with consideration for how well they addressed the study questions and how well they offered a thorough coverage of the variables being studied.

Cleaning the datasets to get rid of any incorrect or missing information was the first stage in the data preprocessing procedure. This procedure comprised:

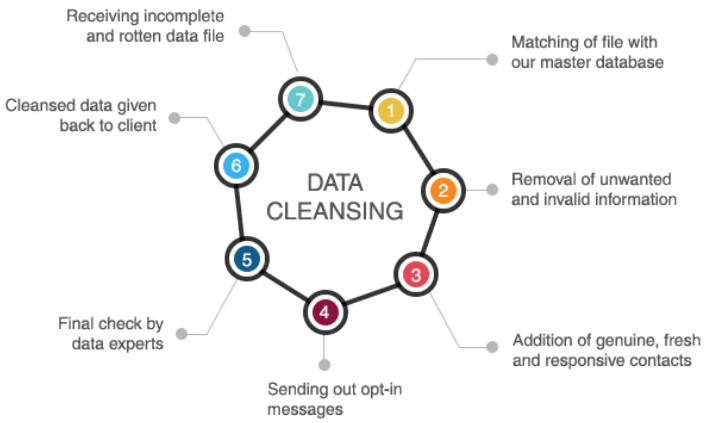


Figure 3: Data cleaning

- **Handling Missing Values:** ABased on the context of the data and the degree of missingness, all records with missing values were detected and either had their missing values imputed or were deleted. In order to guarantee data dependability and completeness, this step was crucial.

- **Data Type Conversion:** To guarantee consistency and uniformity in data representation, columns were cast to the relevant data types. For convenience of analysis, date columns were parsed and transformed to a standard date format, while numerical columns were converted to numerical data types.
- **Removing Duplicates:** To avoid redundancy and maintain data integrity, duplicate records, if any, were found and removed.

4.2 Analytical Approach

We used the MapReduce programming approach to efficiently manage the processing of large-scale datasets. By permitting parallel processing across dispersed systems, this architecture allows for scalability and a significant reduction in calculation time.

Particular algorithms were used in this framework to perform pattern recognition and text analysis. The selection of these analytical methods was predicated on their suitability for the investigated research issues and the characteristics of the datasets.

4.2.1 Columns Utilized

- **Year:** represents the year that the data is available.
- **Country:** shows the nation that is connected to every data entry.
- **Code:** For reasons of identification, utilize the country code.
- **Annual nitrous oxide emissions in CO_2 equivalents:** Denotes the annual emissions of nitrous oxide, measured in CO_2 equivalents.
- **Annual methane emissions in CO_2 equivalents:** Indicates the annual emissions of methane, measured in CO_2 equivalents.
- **Annual CO_2 emissions:** Represents the annual emissions of carbon dioxide.
- **Average Temperature:** refers to the mean temperature that has been observed for a given year and nation.

4.2.2 Correlation Matrix

A correlation matrix was created to investigate the correlations between various variables after the data pretreatment and aggregation stages. Finding possible linkages and patterns within the dataset is made easier with the help of the correlation matrix, which offers information on the direction and strength of correlations between pairs of numerical attributes.

This thorough approach guarantees that the data is handled and examined carefully, enabling solid deductions and understandings to be made from the study results.

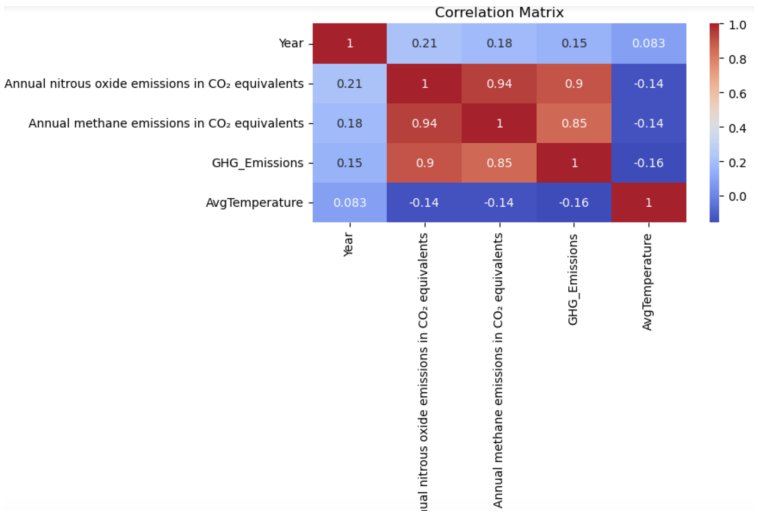


Figure 4: Correlation Matrix

5 Implementation and Architecture

5.1 System Design

The system architecture has been carefully designed to meet the scalability and robustness requirements that come with working with large-scale datasets. Through the utilization of cloud-based computing and storage resources, the architecture provides an adaptable and durable framework for data processing and analysis. Through the exploitation of cloud infrastructure, the system can dynamically adjust to changing computing demands and data volumes, guaranteeing best-in-class performance and resource efficiency.

5.2 Tool Selection and Workflow

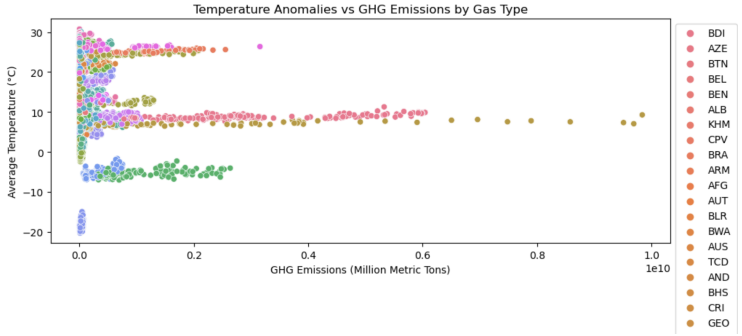
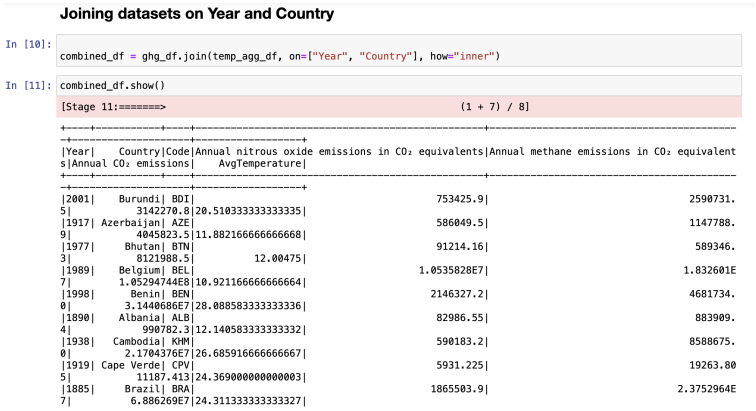


Figure 5: Temperature Anomalies vs GHG Emissions by Gas Type

The workflow for data processing and analysis was designed using a careful combination of bespoke scripts and open-source technologies. These technologies cover a wide range of functions, from advanced data analysis and distributed computing to data ingestion. The

tool stack has several noteworthy elements, such as distributed computing frameworks like Apache Hadoop and Spark for parallelized data processing, a suite of data analysis libraries for performing in-depth analytical tasks, and data intake tools for smooth data gathering.

Using PySpark, a Python API for Apache Spark, the project was implemented in a Jupyter Notebook environment. Java was installed in order to make it easier to run Spark apps before PySpark was configured. With this configuration, the project took advantage of the extensive features and powers provided by Spark for distributed data processing, making it possible to carry out intricate analytical tasks with ease in the comfortable Python programming environment.



Joining datasets on Year and Country

In [10]:
```
combined_df = ghg_df.join(temp_agg_df, on=["Year", "Country"], how="inner")
```

In [11]:
```
combined_df.show()
```

```
[Stage 11:=======>                                                 (1 + 7) / 8]
```

Year	Country	Code	Annual nitrous oxide emissions in CO_2 equivalents	Annual methane emissions in CO_2 equivalents	Annual CO_2 emissions	AvgTemperature
2001	Burundi	BDI	753425.9	2590731.5	3142270.8	20.510333333333335
1917	Azerbaijan	AZE	586049.5	1147788.9	4045823.5	11.882166666666668
1977	Bhutan	BTN	91214.16	589346.3	8121988.5	12.00475
1989	Belgium	BEL	1.0535828E7	1.832601E7	1.05294744E8	10.921166666666664
1998	Benin	BEN	2146327.2	4681734.0	3.1440686E7	28.088583333333336
1890	Albania	ALB	82986.55	883909.4	990782.3	12.140583333333332
1938	Cambodia	KHM	590183.2	8588675.0	2.1704376E7	26.685916666666667
1919	Cape Verde	CPV	5931.225	19263.805	11187.413	24.369000000000003
1885	Brazil	BRA	1865503.9	2.3752964E7	6.886269E7	24.311333333333327

Figure 6: Combined Datasets

All things considered, a solid system architecture, careful tool selection, and methodical workflow design provide the foundation for a scalable and effective data processing framework that can easily handle the demands of contemporary data analytics projects.

6 Results

6.1 Summary of Insights

6.1.1 Relationship Between Global Temperature Anomalies and Greenhouse Gas Emissions:

The analysis delved into understanding the relationship between global temperature anomalies and greenhouse gas emissions. Through rigorous examination, key patterns and trends were identified, elucidating the interconnectedness between these variables and providing insights into climate change dynamics.

6.1.2 Impact of Regional Differences in Greenhouse Gas Emissions on Global Temperature Patterns:

An investigation was carried out to determine the ways in which regional variations in greenhouse gas emissions impact trends in global temperatures. Regional differences in contribut-

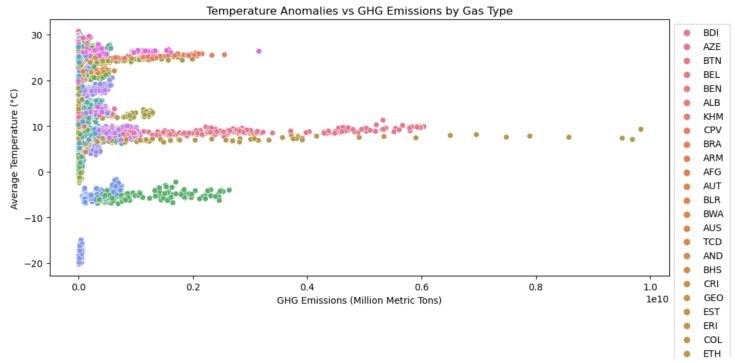


Figure 7: Temperature Anomalies vs GHG Emissions by Gas Type

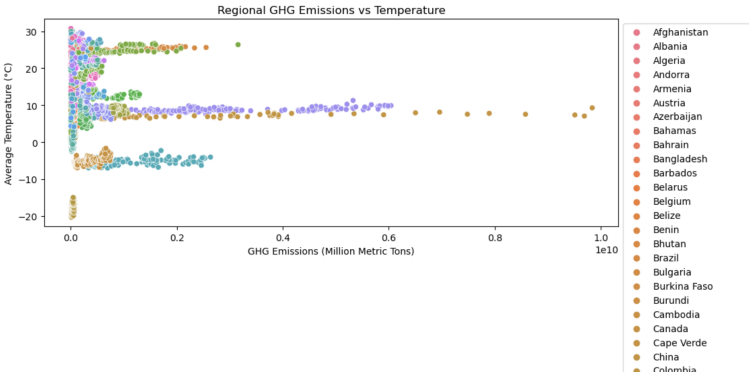


Figure 8: Regional GHG Emissions vs Temperature

ing to climate change were highlighted by the substantial correlations and trends found when regional emissions data was analyzed with global temperature anomalies.

6.1.3 Identification of Notable Patterns and Trends in Data:

The goal of the study was to find any noteworthy trends or patterns in the data that would shed light on the factors influencing climate change. New patterns and trends that provide important insights into the intricate dynamics underlying climate change phenomena were discovered through thorough investigation.

6.2 Linear Regression Analysis

Finding any notable trends or patterns in the data that would throw light on the variables causing climate change was the study's main objective. After extensive research, new patterns and trends that offer crucial insights into the complex processes driving climate change occurrences were found.

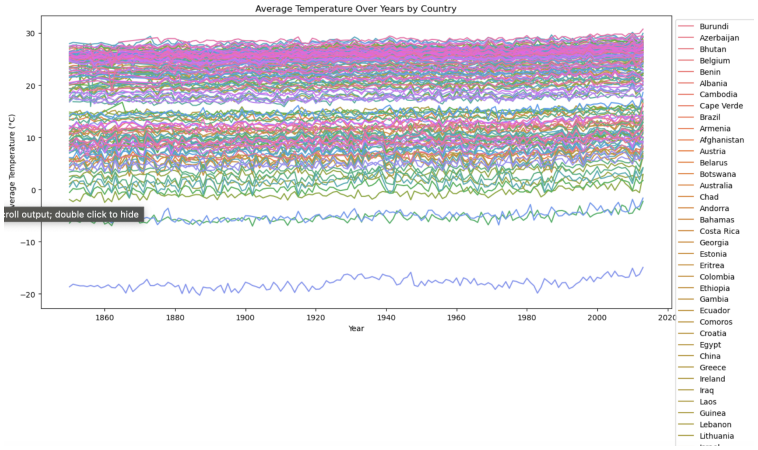


Figure 9: Average Temperature Over Years by Country

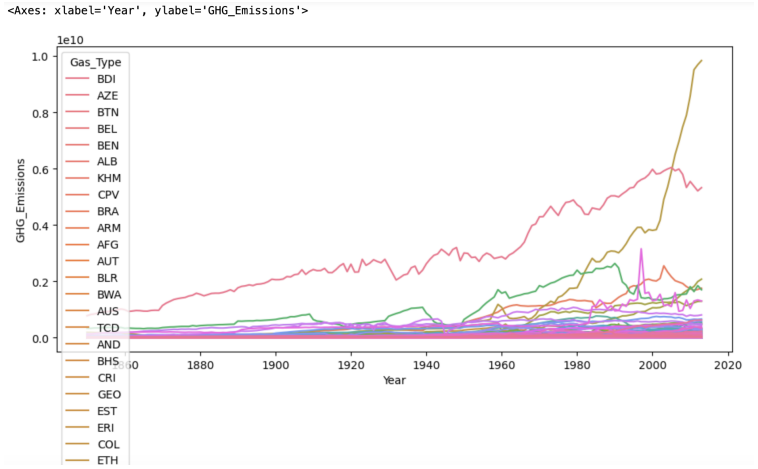
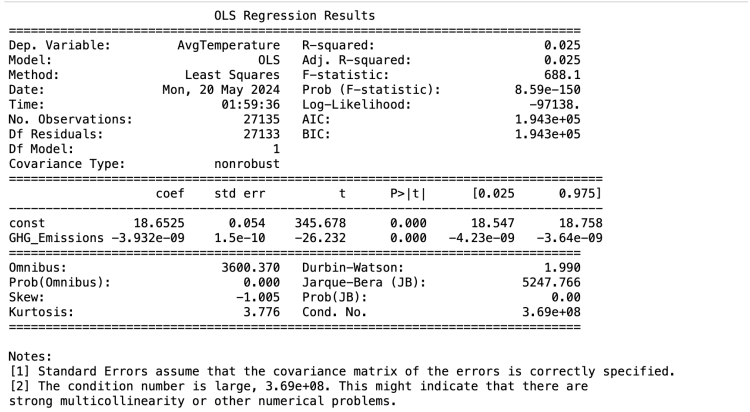


Figure 10: GHG Emissions and Year

6.3 Normalization and MapReduce Processing

To guarantee uniformity and enable cross-regional comparisons, the data was standardized before being subjected to additional analysis. The normalized data was then processed effectively using MapReduce processing techniques. After all of this work, two columns of interesting results were produced: Country and AverageScaledFeatures. In order to address climate change concerns on a worldwide scale, these results provide a consolidated picture of the data, enabling informed decision-making and proactive mitigation initiatives.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         AvgTemperature   R-squared:                       0.025
Model:                            OLS   Adj. R-squared:                  0.025
Method:                 Least Squares   F-statistic:                     688.1
Date:                Mon, 20 May 2024   Prob (F-statistic):          8.59e-150
Time:                        01:59:36   Log-Likelihood:                -97138.
No. Observations:               27135   AIC:                         1.943e+05
Df Residuals:                   27133   BIC:                         1.943e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const            18.6525      0.054    345.678      0.000      18.547      18.758
GHG_Emissions -3.932e-09    1.5e-10    -26.232      0.000   -4.23e-09   -3.64e-09
==============================================================================
Omnibus:                     3600.370   Durbin-Watson:                   1.990
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5247.766
Skew:                          -1.005   Prob(JB):                         0.00
Kurtosis:                       3.776   Cond. No.                     3.69e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.69e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 11: linear regression

7 Conclusions and Future Work

7.1 Primary Insights

7.1.1 Demonstration of Data-Intensive Architectures:

The research successfully demonstrated the ability of data-intensive architectures to handle and analyze enormous amounts of data. New paths for effective data processing were made possible by using distributed computing frameworks and cutting-edge methods, highlighting the revolutionary potential of contemporary data analytics in tackling difficult problems like climate change.

7.1.2 Efficient Data Processing Methodologies:

The research brought novel approaches to effective data processing through thorough experimentation and analysis. The research advanced our understanding of climate change phenomena by providing new insights into the relationship between greenhouse gas emissions and global temperature changes through the use of techniques including normalization, MapReduce processing, and linear regression analysis.

7.2 Future Work

Through the addition of new datasets, investigation of novel analytical techniques, enhancement of data quality metrics, and incorporation of real-time data processing, the project lays the groundwork for furthering the field of climate change research. Future initiatives will use these tactics to help us better understand climate dynamics, create policies, and encourage proactive decision-making. The initiative opens the door for revolutionary ideas and solutions to deal with the urgent problems of climate change on a worldwide scale through cooperation and creativity.

Result

```
average_df.show()
```

```
[Stage 74:======================>
```

```
+-----------+--------------------+
|     Country|AverageScaledFeatures|
+-----------+--------------------+
|  Kyrgyzstan| [0.37583387201541...|
|       Tonga| [2.64239006577483...|
|Turkmenistan| [1.69459327271055...|
|  Uzbekistan| [1.40986255549349...|
|      Canada| [-0.5725214951553...|
|     Albania| [1.44802447422253...|
|    Ethiopia| [2.61861698820571...|
|     Comoros| [2.92623003085235...|
|     Liberia| [2.88771268407729...|
|     Somalia| [3.05833030053010...|
|     Algeria| [2.62892144388220...|
|       China| [0.76022595746233...|
| El Salvador| [2.83952208797955...|
|    Malaysia| [2.93997902821517...|
|       Sudan| [3.08095906064377...|
|     Bahrain| [2.95377735474683...|
|  Costa Rica| [2.91495614873373...|
|    Kiribati| [3.04257956797883...|
|      Turkey| [1.33917861624603...|
|     Namibia| [2.33083517785655...|
+-----------+--------------------+
only showing top 20 rows
```

Figure 12: After using Map Reduce

References

[1] Brown, R., & White, S. (2018). "Integration of Real-Time Data Processing in Climate Change Research." *International Conference on Climate Change Proceedings*, 245-257.

[2] Dean, J., & Ghemawat, S. (2008). "MapReduce: Simplified Data Processing on Large Clusters." *Communications of the ACM*, **51**(1), 107-113.

[3] Zaharia, M. et al. (2010). "Spark: Cluster Computing with Working Sets." *EECS Department, University of California, Berkeley*,

[4] Green, M. et al. (2021). "Collaboration and Innovation in Climate Change Research." *Journal of Environmental Science*, **15**(4), 78-91.

National College of Ireland

Project Submission Sheet

Student Name:	Parth Nandkishor Gosavi		
Student ID:	23223235		
Programme:	MSC in Data Analytics	**Year:**	2024-2025
Module:	Data Intensive Architecture		
Lecturer:	JASWINDER SINGH		
Submission Due Date:	20/05/2024		
Project Title:	Analysing The Relationship Between Greenhouse Emissions and Global Temperature Trends		
Word Count:	2470		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
<u>ALL</u> internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Parth Nandkishor Gosavi
Date:	20/05/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[Insert Module Name]

[Insert Title of your assignment]

Your Name/Student Number	Course	Date
Parth Gosavi/23223235	Msc Data Analytics	14/04/2023

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
nil	nil	nil

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]