National College of Ireland
Project Submission Sheet

Student Name: Utkarsh Satpute, Tushar Gharpure, Pintoo Baghel

Student ID: x23135760, x23289902, x23287501

Programme: MSc. Data Analytics **Year:** 2024 - 25

Module: Data Analytics

Lecturer: Prof. Jaswinder Singh

Submission Due Date: Dec 16, 2024

Project Title: Global Insights into Greenhouse Gases and Air Quality (A Per Capita Perspective)

Word Count: 3423

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.

ALL 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 resultin disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion ab out the validity of their submitted work.

Signature: Utkarsh Satpute, Tushar Gharpure, Pintoo Baghel

Date: 16 / 12 / 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 yourown 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 specificallyrequested.
4. You must ensure that all projects are submitted to your Programme Coordinator onor 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

Global Insights into Greenhouse Gases and Air Quality – Per Capita

Your Name/Student Number	Course	Date
Utkarsh Satpute	MSc. Data Analytics	16 / 12 / 2024
Tushar Gharpure	MSc. Data Analytics	16/ 12 / 2024
Pintoo Baghel	MSc. Data Analytics	16/ 12 / 2024

This section is a supplement to the main assignment, to be used if AI was used in anycapacity 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 Acknowledgement

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

Tool Name	Brief Description	Link to tool
NA	NA	NA

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] **NA**	
[Insert Sample prompt] **NA**	[Insert Sample response] NA

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.

Global Insights into Greenhouse Gases and Air Quality (A per Capita Prespective)

Tushar Gharpure
MSc.in Data Analytics
National College of Ireland
23289902

Pintoo Baghel
MSc.in Data Analytics
National College of Ireland
23287501

Utkarsh Satpute
MSc. In Data Analytics
National College of Ireland
23135760

***Abstract*: This report gives an in-depth analysis of the global greenhouse gas emissions, air pollution, air quality (AQI) levels. It examines how these emissions affect air quality, and their connection with socio-economic factors and the impact they have on public health. We have used datasets spanning from the year 2000 to 2021. We analyzed the dataset to understand and explore movement in emissions, their geographic distribution, and their correlation with the "AQI" levels. This study aims to analyze and provide insights for the leaders, policymakers, scientists, researchers and government to understand and reduce environmental challenges [1].**

***Keywords*—Greenhouse Gas Emissions, Air Quality Index, Environmental Analysis, Global Trends, Data Visualization**

I. INTRODUCTION

In today's day to day life the most challenging concern is climate change, air pollution and causing health effects due to this. Air pollution, greenhouse gas emission from human activities, deforestation, etc. These are the main causes of climate change and health issues. While poor air quality is the leading reason for affecting or causing respiratory health diseases.

In this project we are analyzing the main factors causing low AQI which leads to many health problems. Our project explains the complex relationship between AQI (air quality index) with income per capita of countries. We aimed to analyze the main causes of degrading the air quality and deeply study about a mis concern regarding AQI and income per capita. The uncovered mis conception about this is people do think like if any country is having more income per capita they will have more emissions [2, 3].

Research Questions

How do greenhouse gas emissions vary across countries from 2000 to 2021?
Most of the greenhouse gas emissions are exhausted by industrial activity (factories), energy production and lack of emissions policies. The study will show case trends occurring emissions and factors affecting AQI. As most of the countries have low AQI due to high emissions levels. We are interactively correlating trends with respect to AQI and income.

What is the relationship between per capita emissions and Air Quality Index (AQI)?
The main analysis in our project is to analysis the air quality index of different countries. As we mentioned we are analyzing the trends which are related to income per capita to emissions. The basic misconception is when the income of any country is on the high side, the contribution towards air pollution will be also high as most of the income or revenue is generated by industries, power plants, and traffic. But this is not a clear image. For this, we have made research analysis in this project regarding emission per capita with respect to air quality index.

We performed the **regression test** for the above question: Our regression model effectively explains 97.2% of the variation in AQI value.

Key findings:

PM2.5 and Ozone: The pollutants are the primary drivers of AQI, with a good positive correlation.
Per Capita Emissions: The emission contributes to the pollution, its direct impact on AQI is not that important, it suggests factors like pollutant levels play a more dominant role.
NO2: This pollutant displays an inverse relationship with AQI, it may be due to interaction within the dataset.
In conclusion we can understand that the specific contributions of various pollutants are beyond the per capita emissions, it is important for effectively addressing the air quality issues [4].

Reason for choosing this dataset for this project

This project, which analyzes three datasets greenhouse gas emissions, air quality, and emission is related and relevant in today's global discussion which surrounds the climate change, pollution, and its effects on public health. With increasing awareness to address the environment related challenges. Analyzing emission and its impact on the air quality has become one of trending and crucial areas of study. This project focuses on two decades of data spanning from 2000 to 2021, using three key datasets mentioned earlier. These datasets provide an opportunity to understand and explore how does emission impacts air quality throughout the countries and regions. It helps to give us an overview and identify patterns and correlations that can help to make informed policy decisions. By integrating these three datasets, the project addresses crucial issues related to sustainability such as climate action, health and clean energy. This research contributes to the ongoing advancement of solutions for the climate change and emphasizes its importance in building a sustainable future.

Insights given to government or any organization by analyzing the project

II. RELATED WORKS

This project expands on previous studies and research. They have found a connection between greenhouse gas emissions compared to energy produced and what were the health results. This analysis develops from such works by combining datasets and by implementing new visualization approaches. The consolidation of power plant dataset along with disease numbers or metrics and the weather data allows for a realistic analysis. The combination of "greenhouse gas emissions" data with "air quality indices" (AQI) and per capita trends provides a wide range framework for analysis.

One such work that relates is Crippa et al. (2020) It conducts an extensive analysis of global greenhouse gas emissions using highly detailed datasets. It was published in Nature Communications as one of the articles, their study focused on the significance of detailed and latest emissions data to acknowledge and understand patterns and notify policy successfully. The research breaks down pollution sources into small parts, helps to focus on the large problem and try to fix them [5, 6].

III. METHODOLOGY

Data Acquisition

Dataset 1 – Annual Emission (CSV)

Goal: We have used the dataset on "emissions by country" from the year 2000 to 2022. It serves as an important data source for this analysis. It includes country-based emission data.

Data Source: The dataset is present on Kaggle, a platform for data science and machine learning professionals. Additionally, the dataset used is structured for integration with the data analysis workflows.

Data Format and Use: The structured data (CSV) allows for uninterrupted cleaning and transformation. These libraries are imperative steps in preparing the dataset for visualizing and analysis.

Considerations and Challenges: To make sure that the dataset remains up to date to reflect running trends was a crucial consideration. One of the challenges was handling inconsistency in the name convention of the countries for aligning this dataset with others used in the project. [7].

Dataset 2 – Global Greenhouse Gas Emission (JSON)

Goal: The goal of this dataset is to analyze environmental movements and their impact on public activities and on climate change. This dataset displays country and substance level emissions information.

Data Source: This is the JSON file, and the dataset is present on the KAPSARC - "King Abdullah Petroleum Studies and Research Center" portal, a reliable source climate and energy related data.

Data Format and Use: This dataset was uploaded to the mongoDB, later it was transformed from semi-structure to structured format for ease of use, integration, and exploration. By doing, this it offered filters such as regions, countries, substances like CO_2, CH_4, N_2O etc. These filters are very useful for specific analyses.

Challenges and Considerations:

Data Integration: The dataset had to be structured in such a way that it aligned with the other datasets in the project, for instance country naming convention, formatting etc.

Substance Analysis: This dataset gives a detailed information about the harmful greenhouse gases i.e. CO_2, CH_4, N_2O. If we need to filter, analyze the data specific to substances. We will need accurate handling while doing the preprocessing.

Data Consistency: In the dataset, if the values are missing or there are null values. It could affect the integrity of the project. We need to make sure that not only this dataset but also others have consistency in timeline and acknowledge if there are any missing values. This is important for relevant temporal analysis [8].

Dataset 3 – Global Air Pollution Dataset (CSV)

Goal: This dataset gives details on global air pollution levels, it highlights the pollutants such as PM10, O_3, PM2.5, NO_2, SO_2, and CO. It is critical to understand the effects of these pollutants not on climate, environment and public health across the global.

Data Source: The dataset was collected from the Kaggle website. This is one of the most trusted websites used by data scientists. Since this data is in structured format it is a good option for this project.

Data Format and Use: This dataset is available on the Kaggle website in the CSV format. It is in tabular structure (i.e. rows and columns) this simplifies most of the pre-processing work like standardization, extraction, and summarization.

IV. DATABASE MANAGEMENT

For easy dataset handling we use a hybrid database management strategy with PostgreSQL and MongoDB. The two CSV datasets –"Global Annual Emissions" and "Global Air Pollution" were stored locally in a PostgreSQL database, namely "test", which made use of the powerful relational features for the structured data management and querying. The" Global Greenhouse Gas Emissions" dataset came in a JSON file, which was operated by a MongoDB Atlas cluster named "tushardb".

The JSON data was beneficial, and the system handled it without any issue by using the API for fetching data, loading it and saving it. This approach of the two-database made sure the maximum storage, smooth accessibility, and expandability. This meant a smooth processing of the dataset for the analysis of the project.

V. PROCESS FLOW DIAGRAMS AND DATA CLEANING, INTEGRATION, TRANSFORMATION

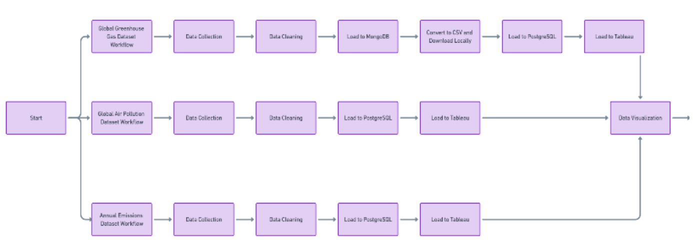


Fig. ETL Flow Chart.

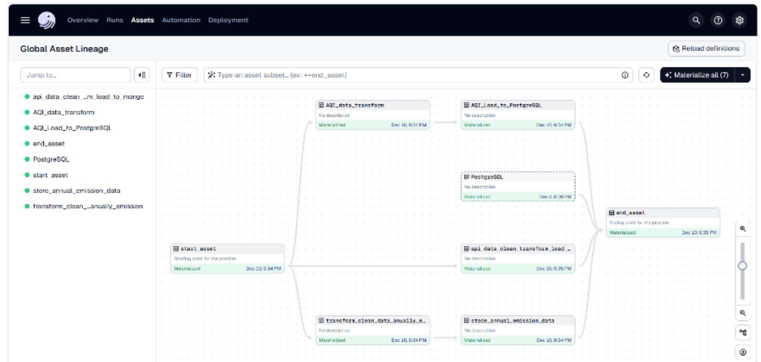


Fig. Dagster Flowchart

Dataset 1: Annual Emissions Dataset Workflow

Data Collection: We obtained the "Annual Emissions" dataset in CSV format from Kaggle.

Data Cleaning: We started inspection for the null values, missing values and formatting errors.

We handled the missing values by either substituting them with averages or medians or discarding the irrelevant rows.

In our project we had to standardize the column names (for e.g., "Country", "Year", "Total Emissions", "Per Capita").

We need to ensure that the data types and numerical columns are consistent.

Database Integration

Step 1: We went on to create a PostgreSQL database named "test".

Step 2: We defined a table schema that is equivalent to the dataset structure (columns such as Country, Year, Total, etc.).

Step 3: We used a Python script to upload the cleaned CSV file into PostgreSQL database.

Data Transformation: We added the calculated columns which were necessary (e.g., per capita emissions trends) and performed the standardization and scaling for stable analysis.

Data Analysis Preparation: We then extracted the cleaned and transformed data from PostgreSQL database and saved processed data back on the system into a local CSV for visualization in Tableau. We started doing the visualization in Tableau then connected to PostgreSQL database to Tableau. We designed the dashboard to visualize the emissions data by country, per capita and the year.

Global Air Pollution Dataset

Data Collection: We Obtained the "Global Air Pollution" dataset in CSV format from Kaggle.

Data Cleaning: We started by removing the null, irrelevant, and missing values (e.g., rows where the city or AQI Value is not present).

We renamed the columns for better clarity (e.g., AQI Value -> AQI_Value, Per_capita -> Per_capita emission).

We validated the data of pollutant-specific AQI categories for consistency.

Database Integration

Step 1: We created a PostgreSQL database named “test”.

Step 2: We defined a table schema with the columns like Country, AQI_Value, CO_AQI_Value, etc.

Step 3: We then uploaded the cleaned CSV file into PostgreSQL using Python scripts.

Data Transformation

We grouped the data by country to calculate average AQI values. We then normalized the AQI categories for consistent country comparisons.

Data Analysis Preparation

We then extracted the transformed data from PostgreSQL and saved it as a CSV for Tableau visualization. We then connected the PostgreSQL database to Tableau.

Global Greenhouse Gas Dataset

Data Cleaning

We parsed the JSON file using Python and converted it into a tabular (CSV) format by using Pandas DataFrame.

We normalized the column names and checked for missing data and handled the anomalies in columns like Emission_Value.

Database Integration

Step 1: We used the MongoDB Atlas to create a cloud database named “tushardb”.

Step 2: We then wrote a Python script to upload the cleaned dataset into MongoDB.

Step 3: We then converted the MongoDB dataset to CSV format for visualization in Tableau.

Data Transformation

We filtered the data for the years 2000–2021 and then calculated the annual averages or totals and added calculated fields like regional emission trends.

Data Analysis Preparation

We exported the transformed MongoDB dataset to a PostgreSQL database for consistency. We saved the processed and cleaned dataset in CSV format for visualization purposes. We connected the PostgreSQL database to Tableau for visualization. We then integrated and merged all the Datasets, for instance merged the "Annual Emissions" and "Global Air Pollution" datasets based on the Country and Year information.

Data Preparation for Analysis

We used Python to clean and standardize the merged data. We generated the calculated fields like Per Capita Emissions, Average AQI, or Total Greenhouse Gas Emissions. We then unified the visualization in Tableau and combined all datasets into a single Tableau dashboard.

Design interactive visuals such as:

A summary table combining emissions and AQI data country wise.

Top 10 countries of high AQI with per capita.

Top 10 countries of low AQI with per capita.

Top countries with high greenhouse value (in Kt) by Per capita.

Top countries with low greenhouse value (in Kt) with Per capita.

VI. DATA SOURCES & DATA DESCRIPTION

Dataset 1: Annual Emissions Dataset

Description: As the name suggests, it provides yearly emissions data on a country-by-country basis, along with the per capita emissions.
Key Columns: country, year, total emissions, per_capita emissions.
Source: Kaggle - Emissions by Country (Quantifying Sources and Emission Levels) [9]

Dataset 2: Global Greenhouse Gas Dataset

Description: This dataset is about tracking greenhouse gases like CO2, CH4, and N2O across various country regions and the substances.
Key Columns: country, year, substance, emission_value.
Source: KAPSARC - “King Abdullah Petroleum Studies and Research Center” Data portal - (Total Global Greenhouse Gas Emissions - This dataset contains Total Emissions by country and main source category from 1970-2021) [10]

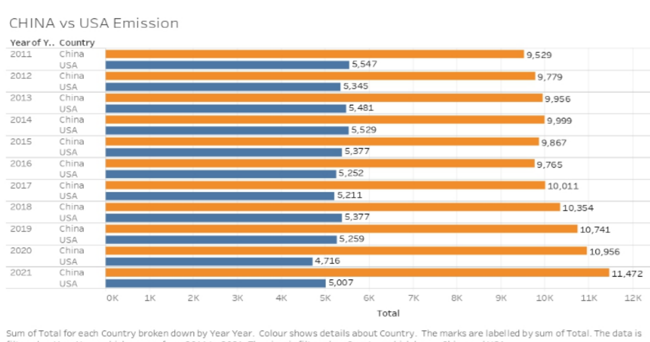
Dataset 3: Global Air Pollution Dataset

Description: This dataset contains the air quality metrics such as “AQI” and specifically substance/pollutant indices.
Key Columns: country, city, aqi_value, pollutant_levels.
Source: Kaggle - This dataset provides geolocated information about the following pollutants: Nitrogen Dioxide (NO2), Ozone (O3), Carbon Monoxide (CO) [11]

VII. CHARTS, PLOTS AND VISUALIZATION

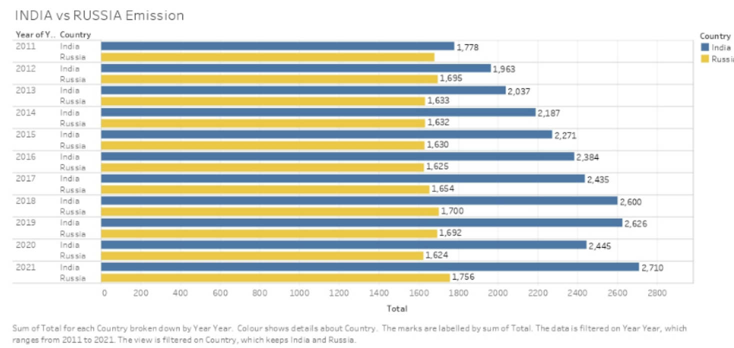
China Vs USA

This graph is a comparison between China and the USA based on greenhouse gas emissions over the period of 10 years. China has been a steady polluter, increasing its emissions to a maximum of 11,472 million tons in 2021. The USA saw a general reduction in emissions, with a small decrease in 2020, and finally reached 5007 million tons in 2021.



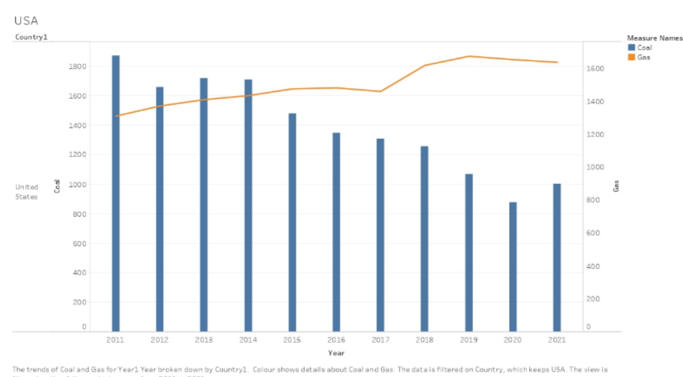
India Vs Russia

This graph indicates the greenhouse gas emissions patterns of India and Russia from 2011 to 2021. India has been on a rising path, with emissions at 1,778 million tons in 2011 to a peak of 2,710 million tons in 2021. In comparison, Russia's average emissions have been almost the same during this time of the decade, moving in the range of 1,600 to 1,700 million tons. More specifically, India succeeded Russia as a leader in emissions in the first years and quietly expanded the lead over time. An extreme cut in emissions for the two countries is the phenomenon to note in 2020, which perhaps is the result of the recent worldwide interruptions.



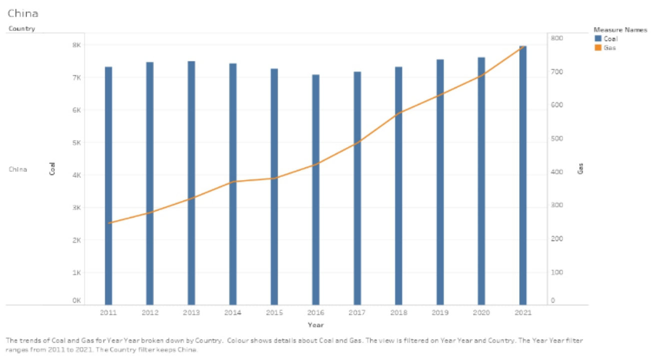
Coal and Gas Emission - USA

The graph shows the trends in the emissions of coal and gas in the USA between 2011 and 2021. The coal emissions show a clear downward trend, which started at about 1,600 million tons in 2011 and led to a substantial cut in 2021. Gas emissions, however, remain relatively steady across the period, and they oscillate between 1,400 and 1,600 million tons. This is evidence of the US's phasing out of coal as a main source of power, with a constant gas use still taking place. The USA's dependency on gas remains on course, despite the decline of coal usage.



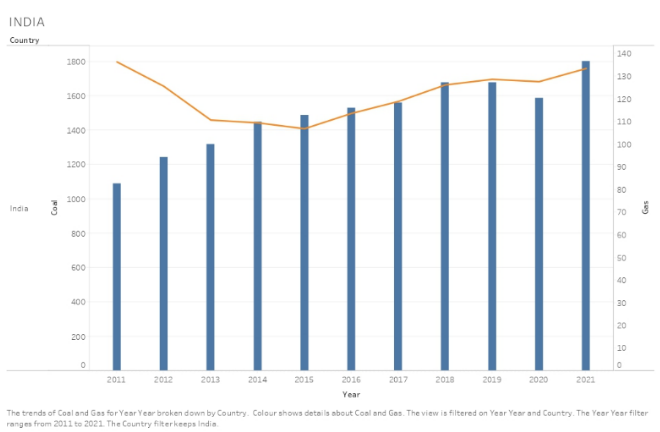
Coal and Gas Usage - China

The diagram shows the tendencies of coal and gas to emit toxic gases to the air in China over the period 2011 to 2021. Coal emissions take the first place, having about 7,000 million tons in 2011 and existing to more than 8,000 million tons by 2021. On the other hand, gas emissions are much lower and lie in a band of 500-800 million tons with a gradual escalation over the years. The trends of both coal and gas emissions are the same thereby, advancing China's industrial activities. The statistics speak of China's horrible usage of coal as the only main energy source.

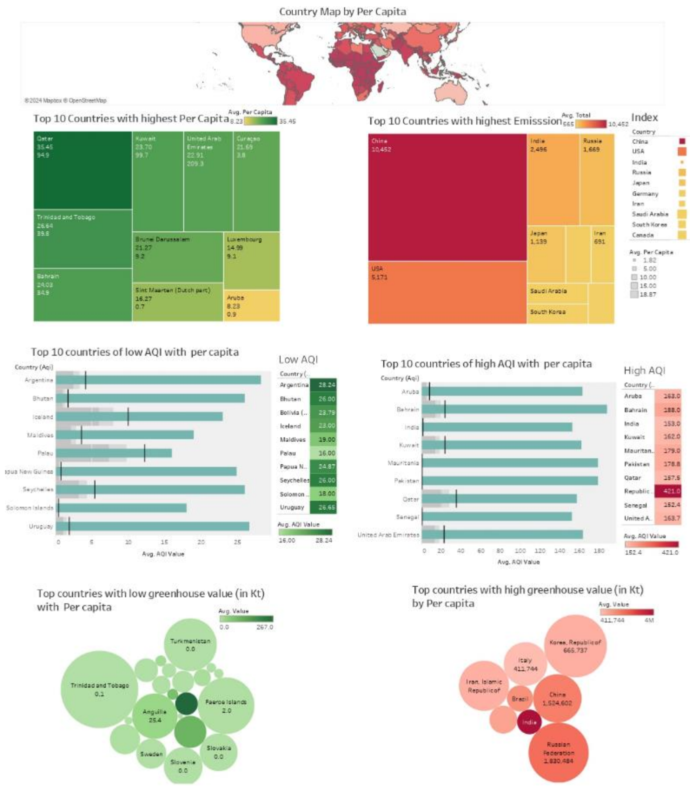


India

This chart below shows that the India’s coal emissions rose steadily from the year 2011 to 2021. The coal releases grew steadily from 1.4 billion tons and reaching almost 1.8 billion tons. This depicts India's dependability on coal for energy and power production. On the other hand, gas emissions have been far lower as they start below 50 million tons and slowly go up to over 100 MT by 2021. Consistent increases in both coal and gas emissions demonstrate the growth of the country's industrial and energy sectors. Broadly, the statistics conveyed India's further engagement with fossil fuels.



Dashboard



Top Emitters:

There are three countries China, USA and India these are the major and primary contributors to the global emission, as they are driven by the industrial production and population. China leads with the highest total of over, 10,452 million tons.

Per Capita Emission: Qatar, Kuwait, and the UAE these countries are fueled by fossil-based economies. They have the highest per capita emission.

Low Emitters: The other countries like Trinidad and Tobago, and Turkmenistan have smaller industry based or have low population, and hence they have lower emissions.

Air Quality Imbalance: There are other nations like Iceland and Argentina which claim that they have extremely clean environment with low AQI, On the other countries like India and Pakistan have significant air pollution challenge.

Regional Trends:The developed countries like Japan, South Korea, even though they have high emissions, still they can maintain moderate AQI levels efficiently managing energy use and controlling the pollution

Extreme Cases: Countries with small population like Aruba and Bahrain. They display both high emissions and poor air quality.

Global Variations: There are major differences in emissions, air quality, and AQI levels. These impacts and highlights diverse energy policies and industrial dynamics [12].

VIII. CONCLUSION:

The project examines data for two decades from 2000 to 2021 to analyze the links between per capita emissions, air quality, and greenhouse gas emissions. Our aim is to understand the complex relationship between these factors and their crucial impact on the environment.

After the analysis that we performed, the result revealed that there are major differences in the greenhouse gas emissions around the countries and regions. The developed nations that have high per capita emissions have over the time displayed that they are better at managing air quality compared to the developing countries, they are still struggling with both degrading air quality and high emissions. A robust correlation between air quality index (AQI) and per capita emission was identified through the linear regression. This showed that there is a desperate need to make the policies stricter for sustainable energy practices for the future.

The datasets that we used were managed and facilitated using the PostgreSQL database for the CSV files and MongoDB Atlas for the JSON file. We incorporated API calls for the data pipeline, it ensured the data accuracy and consistency. We have used Tableau to visualize the data, created the dashboard. It highlights the global trends, country wise variation etc. This project focuses on the significance of combining datasets to get insights and to contribute towards the policy making and environmental awareness [13].

Future Scope

Following could be few of the points that we can consider for future advancement

We can extend the analysis to accommodate the public/healthcare datasets for much deeper insights about the disease correlations for example respiratory disease, cardiovascular conditions etc.

Furthermore, we can employ advanced machine learning models to predict future trends in emission and air quality. These predictions provide crucial information to support informed decision–making.

Our project research can give details to policymakers and governments by providing insights that are actionable, so it is easily translated into practical environmental strategies, such as setting emission reduction goals and improving air quality [14, 15].

Reference

[1] World Health Organization, “Climate change,” World Health Organization, Oct. 12, 2023. Available: https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

[2] “Air pollution vs. GDP per capita,” Our World in Data. Available: https://ourworldindata.org/grapher/air-pollution-vs-gdp-per-capita

[3] D. Zhang, Wullianallur Raghupathi, and Viju Raghupathi, “Exploring the Effects of Greenhouse Gases and Particulate Emissions on Quality of Life: A Country-Level Empirical Study,” Climate, vol. 12, no. 11, pp. 176–176, Nov. 2024, doi: https://doi.org/10.3390/cli12110176. Available: https://www.mdpi.com/2225-1154/12/11/176

[4] National Institute of Environmental Health Sciences, “Human Health Impacts of Climate Change,” National Institute of Environmental Health Sciences, Nov. 07, 2022. Available: https://www.niehs.nih.gov/research/programs/climatechange/health_impacts

[5] M. Crippa et al., “High resolution temporal profiles in the Emissions Database for Global Atmospheric Research,” Scientific Data, vol. 7, no. 1, Apr. 2020, doi: https://doi.org/10.1038/s41597-020-0462-2

[6] M. J. Eckelman and J. D. Sherman, “Estimated Global Disease Burden From US Health Care Sector Greenhouse Gas Emissions,” American Journal of Public Health, vol. 108, no. S2, pp. S120–S122, Apr. 2018, doi: https://doi.org/10.2105/ajph.2017.303846

[7] Wikipedia Contributors, “Regulation of greenhouse gases under the Clean Air Act,” Wikipedia, Jun. 19, 2019. Available: https://en.wikipedia.org/wiki/Regulation_of_greenhouse_gases_under_the_Clean_Air_Act

[8] Richar. T. CARSON, Y. JEON, and D. R. McCUBBIN, “The relationship between air pollution emissions and income: US Data,” Environment and Development Economics, vol. 2, no. 4, pp. 433–450, Nov. 1997, doi: https://doi.org/10.1017/s1355770x97000235

[9] “Emissions by Country,” www.kaggle.com. Available: https://www.kaggle.com/datasets/thedevastator/global-fossil-co2-emissions-by-country-2002-2022

[10] “Total Global Greenhouse Gas Emissions,” Kapsarc.org, 2021. Available: https://datasource.kapsarc.org/explore/dataset/total-global-greenhouse-gas-emissions/information/?disjunctive.world_region&disjunctive.country&disjunctive.substance

[11] “Global Air Pollution Dataset,” www.kaggle.com. Available: https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

[12] R. M. Candanosa, “Reducing Emissions to Lessen Climate Change Would Yield Dramatic Health Benefits by 2030,” Climate Change: Vital Signs of the Planet, Nov. 30, 2021. Available: https://climate.nasa.gov/news/3134/reducing-emissions-to-lessen-climate-change-would-yield-dramatic-health-benefits-by-2030/

[13] Wikipedia Contributors, “Fossil fuel,” Wikipedia, Feb. 07, 2019. Available: https://en.wikipedia.org/wiki/Fossil_fuel

[14] N. Raymond, “Montana urges state top court to overturn landmark climate ruling,” Reuters, Jul. 10, 2024. Available: https://www.reuters.com/legal/montana-ask-state-top-court-overturn-landmark-climate-ruling-2024-07-10/

[15] J. Tauschinski and A. Mooney, “How oil and gas companies disguise their methane emissions,” @FinancialTimes, Nov. 12, 2024. Available: https://www.ft.com/content/374e78dc-69ae-419e-a583-0842b3833d9d