DAS732: Data Visualisation Assignment 3 Report

Ashirwad Mishra
IMT2022108
IIIT-Bangalore
Ashirwad.Mishra@iiitb.ac.in

Krish Patel
IMT2022097
IIIT-Bangalore
Krish.Patel@iiitb.ac.in

Vansh Sinha
IMT2022122
IIIT-Bangalore
Vansh.Sinha@iiitb.ac.in

***Abstract*—This report analyzes GHG emissions using a visual analytics workflow across three tasks: examining industry-level emissions over time, identifying commodity-specific emission drivers, and comparing emission factors across sectors. Industries are analyzed to identify trends and key contributors, leveraging clustering techniques in one iteration. Commodities are evaluated to determine their impact on overall emissions, and sectors are compared for emission factor variability. Machine learning techniques are incorporated in select iterations to enhance analysis and provide deeper insights for emission reduction strategies.**

Introduction

The increasing complexity of data in environmental analysis, particularly in greenhouse gas (GHG) emissions, necessitates sophisticated methods to extract actionable insights. A **visual analytics workflow** is a powerful approach that combines automated data processing, interactive visualizations, and iterative user input, forming a *human-in-the-loop* framework. This methodology enables the synthesis of machine learning techniques with domain expertise, fostering deeper understanding and targeted decision-making.

A visual analytics workflow typically consists of the following components:

- **Data Preprocessing:** Initial stages involve cleaning, normalizing, and integrating datasets to ensure consistency and readiness for analysis.
- **Visualization:** Graphical representations such as time-series trends, comparisons, and multi-dimensional plots allow users to explore data intuitively.
- **Modeling and Analysis:** Incorporation of computational techniques like clustering or regression to identify patterns, predict trends, or classify data.
- **Iterative Feedback Loop:** Users iteratively interact with visualizations, refining parameters, adjusting models, and focusing on areas of interest to uncover actionable insights.

The human-in-the-loop aspect is integral to this workflow, enabling analysts to guide automated processes, validate outcomes, and contextualize findings with domain knowledge. This collaborative interaction not only enhances the quality of insights but also aligns analysis with specific objectives, such as identifying high-emission industries or understanding sectoral variations in GHG emissions.

In this report, the visual analytics workflow is applied iteratively across three tasks—industry-level emissions over time, commodity-specific emission drivers, and sectoral comparisons. By integrating preprocessing, visualization, and modeling with user-driven refinements, the workflow facilitates actionable strategies for emission reduction while emphasizing transparency and adaptability in analysis.

Original Dataset Description

The Original Dataset used in this assignment is named "Supply Chain Greenhouse Gas Emission Factors for US Industries and Commodities" and is publicly available on this website - Dataset. This contains year-wise (2010-2016), Summary and Detailed versions of GHG emissions from Commodities and Industries in the US Economy. The fields/columns in the dataset are as follows:

1. Commodity Code - Code of the commodity or industry from the BEA Make and Use Tables 2012 categorization. 'Detail' and 'summary' are two levels of detail BEA publishes economic input-output accounts data at. 'Detail' level is the most resolved categorization and includes 405 commodity or industry sectors. 'Summary' level is a categorization with medium resolution and includes 73 commodity and 71 industry sectors.

2. Commodity Name - Name of the commodity or industry from the BEA Make and Use Tables 2012 categorization, except detail commodities, which use USEEIO v1.1 names, see Ingwersen and Yang 2017)

3. Substance - Greenhouse gas: 'carbon dioxide' is CO2; 'methane' is CH4; 'nitrous oxide' is N2O; and 'other GHGs' include HFC-23, HFC-32, HFC-125, HFC-134a, HFC-143a, HFC-236fa, CF4, C2F6, C3F8, C4F8, SF6, and NF3

4. Unit - Unit of emission factors for each gas. 'Other GHGs' are aggregated and reported in CO2e (carbon dioxide equivalents) using the IPCC AR4 100-year GWP factors. Purchaser price is the price paid by the consumer and equals to the producer prices plus any associated margin, which generally include distribution, wholesale and retail costs.

5. Supply Chain Emission Factors without Margins - Direct and indirect GHG emissions associated with production of commodity or industry from cradle to the point of production(kg) per 2018 USD of that commodity or industry in the US in purchaser price .

6. Margins of Supply Chain Emission Factors - Direct and indirect GHG emissions associated with production of commodity or industry from the point of production to the point of sale (kg) per 2018 USD of that commodity or industry

in the US in purchaser price of that commodity or industry in the US.

7. Supply Chain Emission Factors with Margins - Direct and indirect GHG emissions associated with production of commodity or industry from cradle to the point of sale(kg) per 2018 USD of that commodity or industry in purchaser price of that commodity or industry in the US.

8. DQ ReliabilityScore of Factors without Margins - Data reliability scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

9. DQ TemporalCorrelation of Factors without Margins - Data temporal correlation scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

10. DQ GeographicalCorrelation of Factors without Margins - Data geographical correlatoin scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

11. DQ TechnologicalCorrelation of Factors without Margins - Data technological correlation scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

12. DQ DataCollection of Factors without Margins - Data collection scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

After cleaning the dataset and creating our dataframe for analysis, we created an additional column : Year - The year in which the data was recorded

Additional Dataset Description

The supplementary datasets (Source - Dataset) used in this analysis, including those for the years 2011, 2014, and 2015, follow a consistent template designed to document greenhouse gas (GHG) emission factors. These datasets enhance the primary *Supply Chain Emission Factors* dataset by providing additional details for different years. Below is a description of one of the tables in the excel sheet. The culmination of several tables creates a single dataset of a particular year. The other tables used will be explained at the point where they are used to reduce the amount of unnecessary content:

- **Fuel Type:** Specifies the category of the fuel, such as coal, coke, fossil fuel-derived solids, or biomass fuels. Subcategories like anthracite coal, bituminous coal, and municipal solid waste further detail the classification.
- **Heating Value (mmBtu per short ton):** Indicates the amount of energy produced per unit of fuel, measured in million British thermal units (mmBtu) per short ton. This value helps assess the energy potential of each fuel type.
- **CO_2 Factor (kg CO_2 per mmBtu):** Represents the quantity of carbon dioxide (CO_2) emissions released per mmBtu of fuel combusted, expressed in kilograms. It is a measure of carbon intensity.
- **CH_4 Factor (g CH_4 per mmBtu):** Provides the amount of methane (CH_4) emitted per mmBtu of fuel combusted, measured in grams. Methane is a potent greenhouse gas with a higher global warming potential than CO_2.
- **N_2O Factor (g N_2O per mmBtu):** Indicates the emissions of nitrous oxide (N_2O) per mmBtu of fuel combusted, measured in grams. Like methane, nitrous oxide has significant climate impacts.
- **CO_2 Factor (kg CO_2 per short ton):** Represents the total carbon dioxide emissions per short ton of fuel combusted, measured in kilograms. This column highlights the absolute emissions potential of each fuel type.
- **CH_4 Factor (g CH_4 per short ton):** Specifies the total methane emissions per short ton of fuel combusted, expressed in grams.
- **N_2O Factor (g N_2O per short ton):** Denotes the total nitrous oxide emissions per short ton of fuel combusted, expressed in grams.

Consistency Across Years

The datasets maintain a uniform structure across different years, ensuring compatibility for longitudinal analysis. This consistency facilitates preprocessing steps such as merging, standardization, and integration with the primary dataset.

Role in Analysis

These additional datasets are crucial for supplementing and validating the primary dataset. They allow for:

- Cross-year comparisons of emission factors to identify trends.
- Standardization of emission factors for uniformity across years.
- Enhanced analysis in visual analytics workflows by providing comprehensive and reliable emission factor data.

This uniform dataset structure enables seamless integration into the visual analytics workflow, supporting iterative analysis and generating actionable insights into GHG emissions.

Workflow Explanation

The provided workflow illustrates the process of analyzing and extracting knowledge from emission data through iterative steps involving data cleaning, visualization, modeling, and subflows for feedback and refinement. Each task leverages several iterations of the below given workflow. Each of the iterations use either of the three SubFlows shown in Fig. 1. Below is an explanation of the components and subflows:

Components of the Workflow

- **Choose Dataset for Model Training:** The workflow begins with selecting a dataset for model training. This dataset contains the primary emission data used for analysis.
- **Emission Data and Additional Data:** These are the primary sources of information. *Emission data* represents the core dataset, while *additional data* complements it by providing supplementary information.
- **Cleaning Process:** The raw data (both emission and additional) undergoes a cleaning process to ensure consistency, completeness, and readiness for analysis. This

process removes invalid or missing values, standardizes formats, and prepares the datasets for downstream tasks.

- **Clean Emission Data and Clean Additional Data:** After the cleaning process, the emission and additional datasets are stored separately as cleaned datasets. These cleaned datasets are critical for ensuring accurate analysis and visualization.
- **Visualizations:** Cleaned datasets are transformed into visual representations, facilitating the mapping of data insights. These visualizations highlight patterns, trends, and anomalies in the emission data and serve as a precursor for further modeling.
- **Model:** This component involves the application of machine learning or statistical models on the cleaned data. Models are trained to extract inferences, predict trends, and provide actionable insights.
- **Knowledge:** The end goal of the workflow is to convert data into knowledge. Insights derived from visualizations and models are aggregated, validated, and presented as meaningful knowledge to stakeholders.

Subflows and Feedback Loops

- **Subflow-1:** This subflow connects visualizations to the knowledge component. It includes a feedback loop, allowing users to refine visualizations based on insights gained from the knowledge. Subflow-1 emphasizes the iterative improvement of visual representations to enhance data interpretation.
- **Subflow-2:** This subflow links the cleaned additional data to the visualization and knowledge components. It introduces a feedback loop for refining the integration of supplementary information into the visualizations and insights. Subflow-2 focuses on ensuring that additional data enriches the primary analysis.
- **Subflow-3:** Subflow-3 represents an iteration of the workflow where machine learning (ML) models are applied to the data. This subflow starts by selecting a dataset—either the original emission data or the additional data—depending on the specific analysis requirements. The chosen dataset is processed through the cleaning and preparation steps, followed by training an ML model to extract insights.

Key Differences Between Subflows

- **Subflow-1** is centered around refining visualizations for better communication of insights.
- **Subflow-2** focuses on integrating additional data into the analysis pipeline to enhance context and depth.
- **Subflow-3** loops back to the cleaning stage, emphasizing data quality improvement informed by insights.

This workflow exemplifies a human-in-the-loop approach, where feedback at various stages ensures iterative refinement.

Tasks Overview

This report focuses on analyzing greenhouse gas (GHG) emissions across industries, commodities, and sectors through the following tasks:

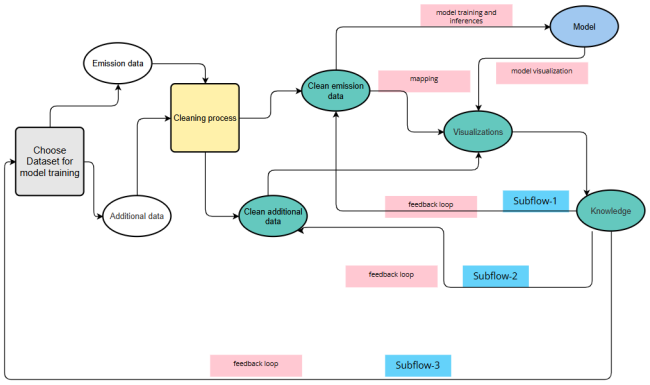


Fig. 1. Workflow illustrating the iterative process of data cleaning, visualization, modeling, and feedback-driven knowledge generation. Subflows highlight the iterative refinement of visualizations, integration of additional data, and application of ML models for insight generation.

- **Task 1: Industry-Level Emissions Over Time** Analyze historical GHG emissions for different industries to identify trends, key contributors, and changes over time.
- **Task 2: Commodity-Specific Emissions and Their Drivers** Examine the relationship between commodity sales and emissions to identify high-emission products and their primary drivers.
- **Task 3: Deeper Analysis into Specific Sectors' Emissions** Conduct an in-depth evaluation of emissions within specific sectors to explore their variability and identify high-emission sub-sectors.

I. Task 1: Industry-Level Emissions Over Time

This section will analyze supply chain emissions of different greenhouse gases (GHGs) across different industries, over the years. We will also deep dive into industry specific data for particular years. We will majorly focus on 5 industries in this particular section, i.e, Utilities, Farms, Petroleum and coal products, Water transportation and Air transportation.

Later in this section we will also analyze the supplementary dataset and create a clustering plot to show the clusters of data of different industries.

All the plots in this section are interactive and thus, we have provided all the plots in html files as well as png format.

Let us start by simple line charts, representing the trends of individual industries GHGs emissions over the years. These plots serve as time series plots depicting the trends and changes in the emission values of different gases for different industries over the years.

Fig.2 provides an overview of greenhouse gas emissions across three industries—Utilities, Farms, and Petroleum and Coal Products—between 2010 and 2016. The emission factors are presented in terms of kilograms of greenhouse gases per 2018-adjusted USD. For the Utilities sector, carbon dioxide is the most significant contributor to emissions, showing a decreasing trend over the years with a slight plateau between 2015 and 2016. Emissions from methane, nitrous oxide, and other greenhouse gases are negligible in comparison. Farms

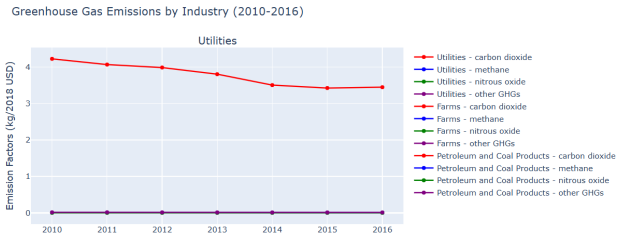


Fig. 2. Time series line chart for Utilities industry

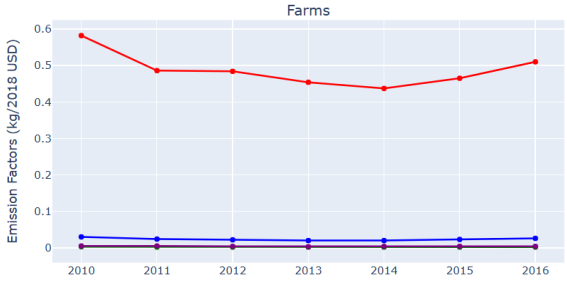


Fig. 3. Time series line chart for Farms industry

demonstrate relatively low emission factors, with carbon dioxide again leading but at much lower levels than the Utilities sector. Petroleum and Coal Products have the highest emissions among the three industries, primarily dominated by carbon dioxide.

Fig.3 focuses on the Farms sector, showcasing carbon dioxide as the primary contributor to emissions, followed by methane and other gases, though these are nearly negligible. The emissions of carbon dioxide exhibit a slight decline from 2010 to 2013, followed by an upward trend from 2014 to 2016. This suggests that despite improvements in emission reduction techniques or practices early on, the industry faced challenges or demand increases that resulted in higher emissions in later years. Methane and other greenhouse gases remain stable and minimal, indicating limited change in these specific emission sources over the observed period.

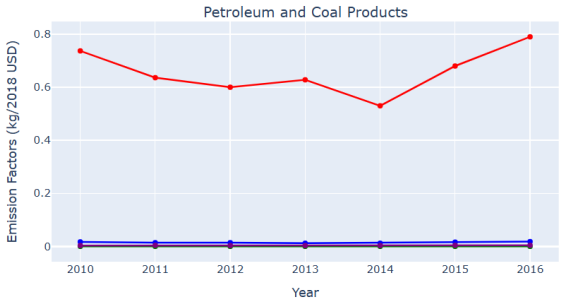


Fig. 4. Time series line chart for Petroleum and coal products industry

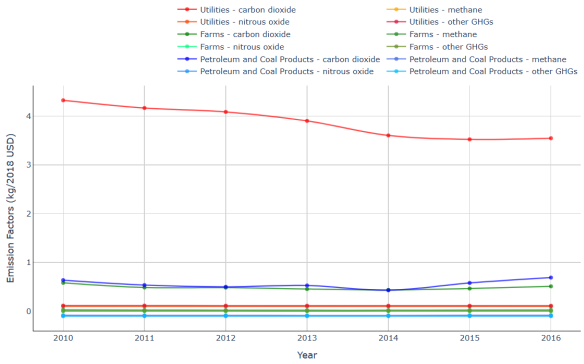


Fig. 5. Time series Braided plot for different industries

Fig.4 highlights the trends in the Petroleum and Coal Products sector, where carbon dioxide emissions dominate with a clear decreasing trend from 2010 to 2014, followed by a noticeable rise from 2014 to 2016. This rebound in emissions could be associated with increased production or a shift in energy policies or market dynamics during those years. Methane and other greenhouse gases remain negligible across all years, indicating that the majority of efforts and challenges in emissions control for this sector revolve around carbon dioxide.

Together, these plots illustrate the varying contributions of industries to greenhouse gas emissions and underscore the critical need for targeted strategies to mitigate emissions in high-contributing sectors like Utilities and Petroleum and Coal Products.

We will now see a braided time series plot for the emissions of gases for different industries.

The image shows a braided line graph that displays the emission factors (in kg/2018 USD) for various greenhouse gas (GHG) sources over the years 2010 to 2016. The lines represent different sectors, including Utilities, Farms, and Petroleum and Coal Products, further broken down by the types of GHGs emitted (carbon dioxide, methane, nitrous oxide, and other GHGs).

The overall trend shows a general decrease in emission factors for most sectors and GHG types over the 7-year period. This suggests that the emissions intensity (emissions per unit of economic output) has been declining, likely due to improvements in technology, efficiency, and/or changes in the energy mix.

A few key observations from the line charts and braided plot:

1. Emissions from Utilities (particularly carbon dioxide and nitrous oxide) show the sharpest declines over the time period.

2. Emissions from Farms also exhibit downward trends, especially for methane and nitrous oxide.

3. Petroleum and Coal Products have relatively stable emission factors, with only slight decreases in carbon dioxide and other GHGs.

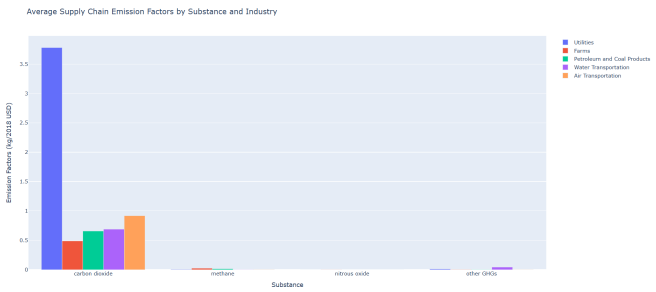


Fig. 6. Bar chart representing average emission values

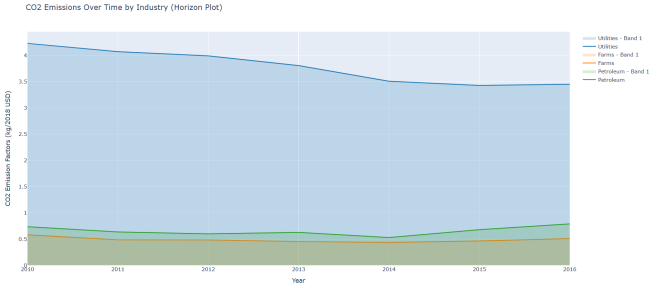


Fig. 7. CO2 Emissions Over Time by Industry (Horizon Plot)

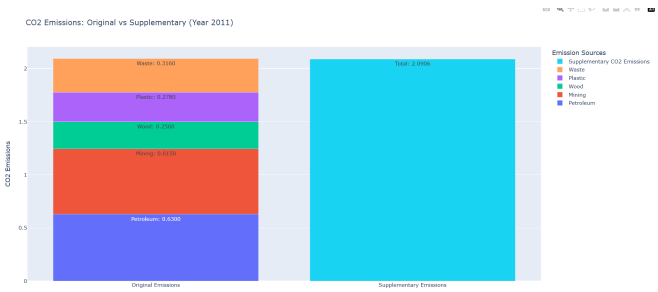


Fig. 8. CO2 Emissions: Original vs Supplementary (Year 2011)

4. The lines for different GHG types within each sector are generally well-differentiated, indicating that the data provides a detailed breakdown of the emissions profile.

The line graph format was likely chosen to effectively visualize the trends over time for the various sectors and GHG types. This type of plot allows for easy comparison of the relative magnitudes and changes in emission factors across the different categories. The use of color-coding and labeling helps to clearly distinguish the various data series.

Overall, this plot provides a comprehensive overview of the emission factors for key economic sectors and GHG types, highlighting the progress made in reducing the greenhouse gas intensity of economic activities over the 2010-2016 period.

”Average Supply Chain Emission Factors by Substance and Industry,” is a bar chart that compares the average supply chain emission factors for different substances and industries. This visualization helps understand the relative environmental impact of various supply chain components. The plot shows that the utilities industry has the highest emission factors for carbon dioxide, methane, and nitrous oxide, indicating it has the greatest environmental impact among the industries shown. The farms and petroleum and coal products industries also have significant emission factors, particularly for carbon dioxide and methane.

The choice of a bar chart for this visualization is appropriate, as it allows for a clear comparison of the emission factors across different substances and industries. The use of distinct colors for each industry and substance makes the data easy to interpret and differentiate. Bar charts are effective at displaying and comparing discrete values, making them well-suited for this type of data. The bars represent the average emission factors for each substance and industry, providing a high-level overview of the relative environmental impact. This information is useful for identifying the industries and substances that require the most attention and prioritization in terms of emission reduction efforts.

The second image, ”CO2 Emissions Over Time by Industry (Horizon Plot),” displays the trend of CO2 emissions over time for different industries. This type of plot, known as a horizon plot, is effective in visualizing multi-variate time-series data, as it allows for the comparison of multiple industries simultaneously while preserving the ability to see detailed trends.

The plot shows a clear downward trend in CO2 emissions for the utilities industry, both in the ”Utilities - Band 1” and ”Utilities” categories. The farms and petroleum industries also exhibit declining emissions, though at a slower rate. The use of a horizon plot enables the viewer to easily identify the overall trends and relative differences between the industries.

The choice of a horizon plot is well-suited for this type of data, as it provides a concise and effective way to visualize the complex relationships between multiple time-series variables. Horizon plots are particularly useful when dealing with a large number of time-series lines, as they avoid the clutter and overlapping that can occur in traditional line charts.

In this case, the horizon plot clearly highlights the diverging trends between the utilities industry and the other industries, making it easy to identify the industries that have been most successful in reducing their CO2 emissions over time. This information can inform decision-making and guide future emissions reduction strategies.

The third image, ”CO2 Emissions: Original vs Supplementary (Year 2011),” presents a comparison of the original and supplementary CO2 emissions for different supply chain components in the year 2011, using the original and supplementary datasets.

The stacked bar chart clearly shows the relative contributions of each supply chain component to the total CO2 emissions, with waste, plastic, wood, mining, and petroleum being the primary contributors. The use of stacked bars allows for easy comparison of the original and supplementary emissions for each component.

The choice of a stacked bar chart is appropriate for this type of data, as it effectively communicates the proportional

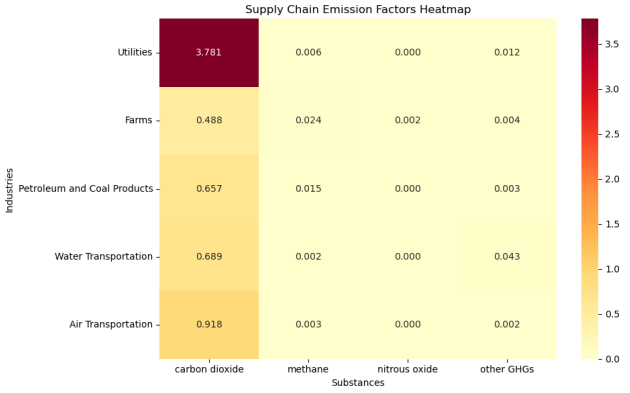


Fig. 9. Supply Chain Emission Factors Heatmap

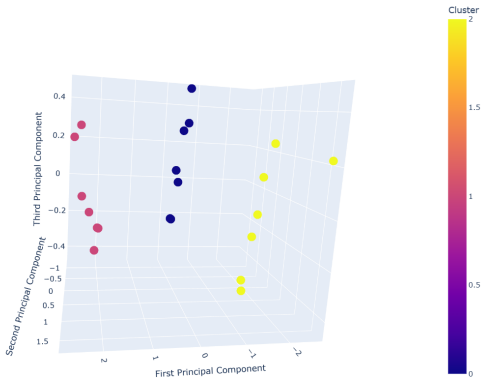


Fig. 10. 3D Scatter plot for K Means clustering

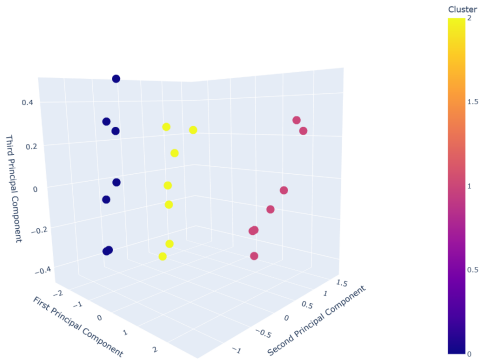


Fig. 11. 3D Scatter plot for K Means clustering

differences between the original and supplementary emissions for each supply chain component. Stacked bar charts are useful when you want to show the breakdown of a total value into its component parts, which is exactly what is being done in this visualization.

By separating the original and supplementary emissions, the chart provides insights into the relative importance of each supply chain component in contributing to the overall emissions. This information can be used to identify the areas where emissions reduction efforts should be focused, such as addressing the primary contributors like waste, plastic, and mining.

The fourth image, "Supply Chain Emission Factors Heatmap," provides a more detailed view of the supply chain emission factors for different substances and industries. The heatmap visualization allows for easy identification of the industries and substances with the highest emission factors, with the utilities industry and carbon dioxide standing out as the most significant contributors.

The choice of a heatmap is well-suited for this type of data, as it enables the viewer to quickly identify patterns and outliers in the emission factors across different industries and substances. Heatmaps are effective at displaying large amounts of data in a compact and visually intuitive way, allowing for easy comparison and identification of high and low values.

In this case, the heatmap clearly highlights the industries and substances with the most significant environmental impact, providing a valuable tool for prioritizing emissions reduction efforts. The use of a color scale, with darker shades representing higher values, further enhances the readability and interpretation of the data.

Overall, the selection of visualization types, including bar charts, horizon plots, stacked bar charts, and heatmaps, effectively communicates the complex relationships and trends within the supply chain emission data. Each plot serves a specific purpose and provides valuable insights that contribute to a comprehensive understanding of the environmental impact of the industries and supply chain components.

Fig 10 shows interactive 3D scatter plot using Plotly Express to visualize the results of the K-Means clustering performed on the emissions data. This type of visualization is well-suited for this analysis as it allows the user to explore the high-dimensional data in an intuitive and interactive manner.

The plot shows the distribution of data points (representing the emissions profiles of different industries and years) across the three-dimensional space defined by the first three principal components. The data points are colored based on their assigned cluster membership, revealing distinct groupings of similar emissions profiles.

The clustering algorithm has identified multiple clusters within the data, indicating that there are significant differences in the emissions characteristics across the industries and time periods represented. By examining the changes in cluster assignments and the relative positions of the clusters in the 3D space, analysts can identify trends and patterns in the evolution of emissions over time and across different industry sectors.

The choice of a 3D scatter plot is appropriate for this type of analysis, as it allows the visualization of the high-dimensional data in a compact and interpretable manner. The use of Plotly Express provides an interactive and customizable plotting interface, enabling the user to rotate, zoom, and hover over data points to obtain additional information, such as the corresponding industry and year.

The plot is further enhanced by the use of a color-coding

```
Cluster Analysis:

Cluster 0:
Industries: ['Petroleum and coal products']
Years: [2010 2011 2012 2013 2014 2015 2016]

Average Emission Factors:
Substance
carbon dioxide    0.657286
methane           0.014714
nitrous oxide     0.000000
other GHGs        0.003429
dtype: float64

Cluster 1:
Industries: ['Utilities']
Years: [2010 2011 2012 2013 2014 2015 2016]

Average Emission Factors:
Substance
carbon dioxide    3.780571
methane           0.006286
nitrous oxide     0.000000
other GHGs        0.012429
dtype: float64

Cluster 2:
Industries: ['Farms']
Years: [2010 2011 2012 2013 2014 2015 2016]

Average Emission Factors:
Substance
carbon dioxide    0.488286
methane           0.023571
nitrous oxide     0.002143
other GHGs        0.004143
```

Fig. 12. Cluster Analysis

scheme to represent the different cluster assignments, making it easy to visually distinguish the groupings within the data. The axis labels and titles provide clear context for interpreting the plot, and the ability to export the visualization as an HTML file allows for easy sharing and further exploration.

The 3D scatter plot is a well-established technique for visualizing the results of dimensionality reduction and clustering algorithms, such as PCA and K-Means. By projecting the high-dimensional data onto the first three principal components, the script is able to capture a significant portion of the variance in the original data, while still maintaining a visually interpretable representation.

The interactive nature of the plot, enabled by Plotly Express, is particularly valuable in this context, as it allows the user to explore the data in a more dynamic and engaging manner. The ability to hover over data points and view the corresponding industry and year information can provide valuable insights and facilitate further investigation of the clustering results.

Overall, the choice of the 3D scatter plot is well-justified, as it effectively communicates the key findings of the clustering analysis and provides a powerful tool for exploring the complex relationships and patterns within the emissions data.

II. Task 2: Commodity-Specific Emissions and Their Drivers

This section will analyze supply chain emissions of different greenhouse gases (GHGs), such as carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and other GHGs, to identify the major contributors and their changes in distribution across sectors and commodities over time. Using iterative visualizations, such as sunburst charts and treemaps, in combination with trend analyses, the objectives here are to identify key actors, examples of significant change, and sector-specific dynamics that include sudden shifts as well as more steady trends in emissions.

It highlights the role of sectors like Transportation, Extractives, Agriculture, Machinery, and Primary Metals in contributing to overall emissions, offering real-world insights into efficient mitigation strategies and policy-making. The visualizations, refined continuously using Subflow-1 in our visual analytics system, provide a more comprehensive analysis by combining emission categories and understanding dominant emission sources and trends. This approach enables a structured and effective strategy for reducing the environmental impact of supply chains.

Iteration 1 Subflow 1 (Contribution)

Carbon Dioxide Emissions—Initial Analysis

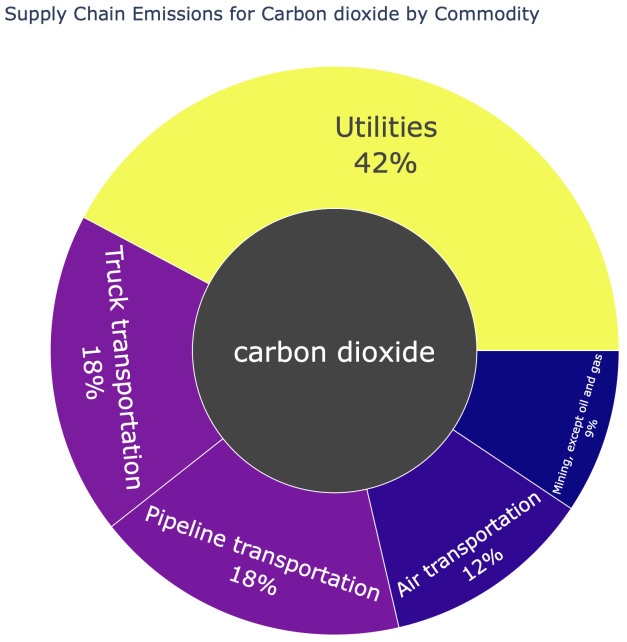


Fig. 13. Carbon Dioxide emissions across different commodities

Context:
Carbon dioxide emissions are a significant contributor to global climate change, and understanding the sources within

supply chains is critical for devising effective mitigation strategies. This visualization sheds light on the relative contributions of various commodities to the carbon dioxide footprint in the supply chain.

Initial Insights from the Sunburst Chart:
The sunburst chart (Fig. 13) provides a breakdown of carbon dioxide (CO_2) emissions across different commodities:

- **Utilities:** Dominate the chart with 42%, highlighting the significant role of energy production and consumption in supply chain emissions.
- **Truck Transportation and Pipeline Transportation:** Each contribute 18%, underscoring the emissions from overland and underground logistics.
- **Air Transportation:** Accounts for 12%, reflecting the impact of rapid logistics.
- **Mining:** Contributes 9%, emphasizing the emissions from resource extraction.

This visualization initially points to utilities as the largest contributor to CO_2 emissions.

Methane Emissions—Initial Analysis

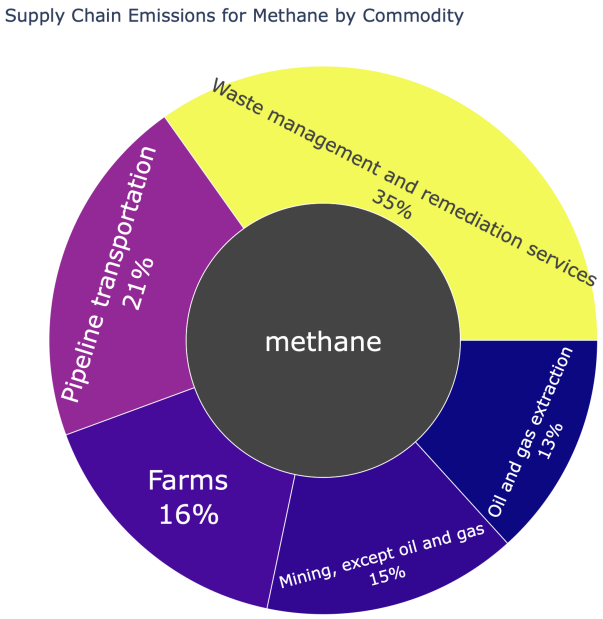


Fig. 14. Methane emissions across different commodities

Context:
Methane emissions are a potent driver of global warming, with a warming potential significantly higher than carbon dioxide over shorter timeframes. Identifying the major contributors within supply chains is essential for developing targeted reduction strategies. This visualization reveals the distribution of methane emissions across key commodities, providing insights into areas with the highest mitigation potential.

Initial Insights from the Sunburst Chart:
The sunburst visualization of methane (CH_4) (Fig. 14) emissions provides an initial breakdown of the key contributors to methane emissions across the supply chain:

- **Waste Management and Remediation Services:** Lead, contributing 35% of methane emissions, underlining the significant impact of organic waste decomposition.
- **Pipeline Transportation:** Accounts for 21%, driven by methane leaks and emissions during natural gas transport.
- **Farms:** Contribute 16%, highlighting the agricultural sector's role, particularly livestock emissions.
- **Mining:** Contributes 15%, emphasizing methane emissions from resource extraction.
- **Oil and Gas Extraction:** Accounts for 13%, underscoring methane emissions from fossil fuel production.

This visualization initially points to waste management as the largest contributor to (CH_4) emissions.

Nitrous Oxide Emissions—Initial Analysis

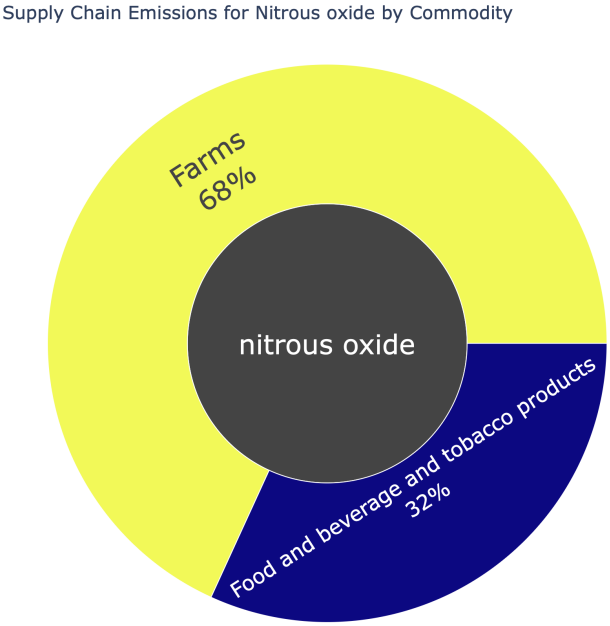


Fig. 15. Nitrous Oxide emissions across different commodities

Context:
Nitrous oxide is a powerful greenhouse gas with long-lasting effects on the atmosphere, contributing to both global warming and ozone layer depletion. Understanding its sources in supply chains is critical for developing sustainable practices. This visualization highlights the relative contributions of different commodities to nitrous oxide emissions, offering a roadmap for effective intervention.

Initial Insights from the Sunburst Chart:
The sunburst chart (Fig. **??**) provides an overview of nitrous oxide (N_2O) emissions in the supply chain, revealing the following:

- **Farms:** Dominate with 68% of total emissions, highlighting the significant role of agricultural practices such as fertilizer use and soil management.
- **Food, Beverage, and Tobacco Products:** Contribute 32%, reflecting emissions from processing, production, and supply chain logistics in this sector.

This initial breakdown directs attention toward agriculture as the most critical area for intervention.

Other GHG Emissions—Initial Analysis

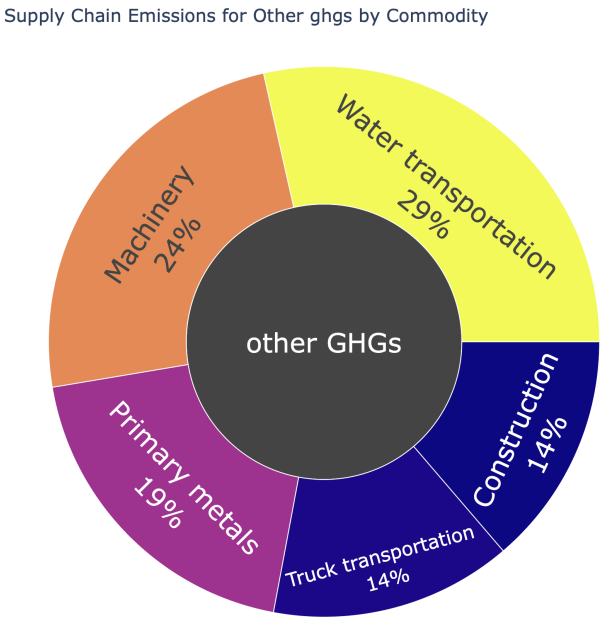


Fig. 16. Other GHG emissions across different commodities

Context:
Beyond carbon dioxide, methane, and nitrous oxide, other greenhouse gases (GHGs) also play a critical role in driving climate change. These gases often originate from specific industrial, transportation, and manufacturing processes. This visualization explores the distribution of these other GHGs within supply chains, emphasizing the importance of addressing emissions from less obvious but impactful sources.

Initial Insights from the Sunburst Chart:
The sunburst chart (Fig. 16) for other greenhouse gases (GHGs) identifies the following contributors:

- **Water Transportation:** Accounts for 29% of total emissions, highlighting the significant impact of shipping and logistics on GHG levels.
- **Machinery:** Contributes 24%, reflecting emissions from manufacturing and operational activities in various industries.
- **Primary Metals:** Account for 19%, underscoring emissions from metal production processes.
- **Truck Transportation:** Contributes 14%, emphasizing emissions from overland logistics.
- **Construction:** Accounts for 14%, highlighting emissions from building activities and related supply chains.

While water transportation emerges as the largest individual contributor, other sectors also play substantial roles in the overall emissions.

Data Processing

I performed data cleaning and processing steps to consolidate related categories and calculate their aggregated values. This involved grouping subcategories, such as different types of transportation or extractive industries, under broader umbrella terms like "Transportation" and "Extractives." Redundant or less significant categories were merged to ensure clarity and focus on the most impactful contributors. The data was restructured to sum the emission values for these consolidated groups, ensuring accurate representation in the treemap. Additionally, percentage contributions were recalculated to reflect the new groupings, providing a refined and simplified view of the data.

Iteration 2 Subflow 1 (Contribution)

Carbon Dioxide Emissions—Refined Analysis

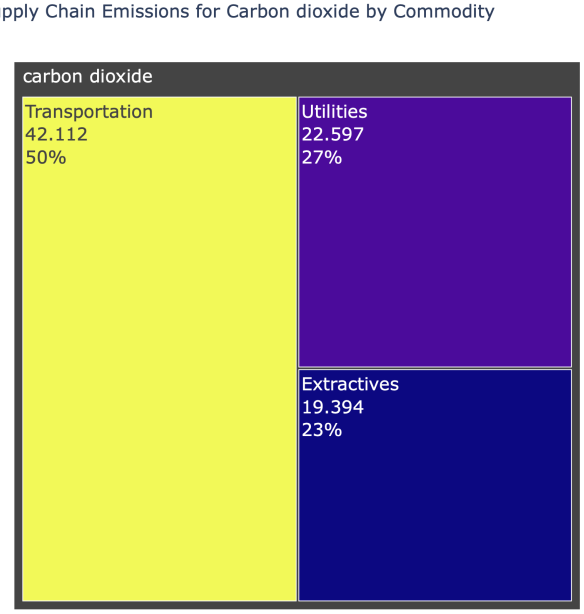


Fig. 17. Carbon Dioxide emissions across different commodities

Refined Insights from the Treemap
The treemap (Fig. 17) consolidates these categories into broader sectors, providing a refined understanding:

- **Transportation:** Emerges as the largest category, contributing 50% (42.112) of total emissions. This consolidates truck, pipeline, and air transportation, highlighting the logistics sector's significant footprint.
- **Utilities:** Account for 27%, reaffirming their substantial impact but shifting them to the second position in importance.
- **Extractives:** Contribute 23%, reflecting emissions from resource extraction and processing activities.

The treemap underscores transportation's dominance, changing the narrative from the initial insights.

Revised Narrative
The transition from the sunburst to the treemap reveals a

key insight: while utilities appear as the largest individual contributor, transportation becomes the leading sector when categories are consolidated. This highlights the need for targeted decarbonization efforts in logistics and supply chain operations.

Methane Emissions—Refined Analysis

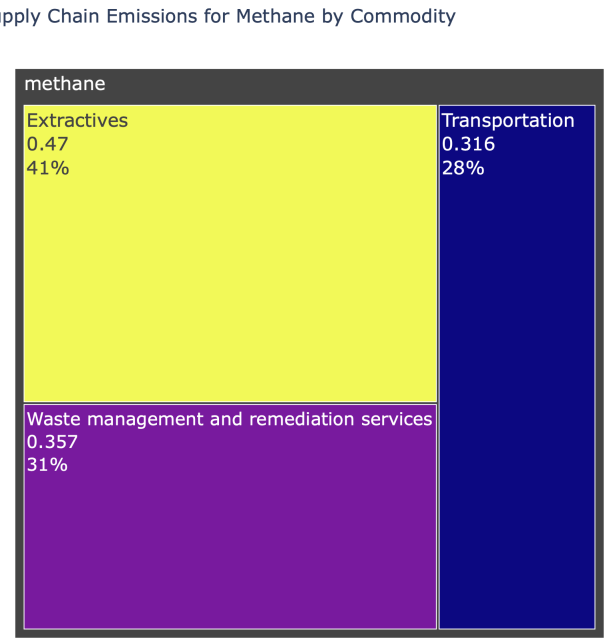


Fig. 18. Methane emissions across different commodities

Refined Insights from the Treemap

After consolidating categories into broader groups, the treemap (Fig. 18) reveals a more accurate perspective of methane (CH_4) emissions:

- **Extractives (41%):** This includes mining, oil, and gas activities, emphasizing the dominant role of the energy sector in methane emissions.
- **Waste Management (31%):** Consolidating waste handling and remediation services, this category highlights the challenges of managing decomposing organic materials effectively.
- **Transportation (28%):** Methane leaks from pipeline transportation emerge as a substantial contributor.

Revised Narrative

The refined treemap shifts our understanding of methane emissions, with extractive industries now identified as the top contributor (41%)—a stark revelation given the previously highlighted dominance of waste management in the sunburst chart. This refinement reinforces the importance of prioritizing emission control strategies in fossil fuel extraction and transportation.

Nitrous Oxide Emissions—Refined Analysis

Refined Insights from the Treemap

The treemap (Fig. **??**) consolidates emission categories to provide a clearer picture of the contributions:

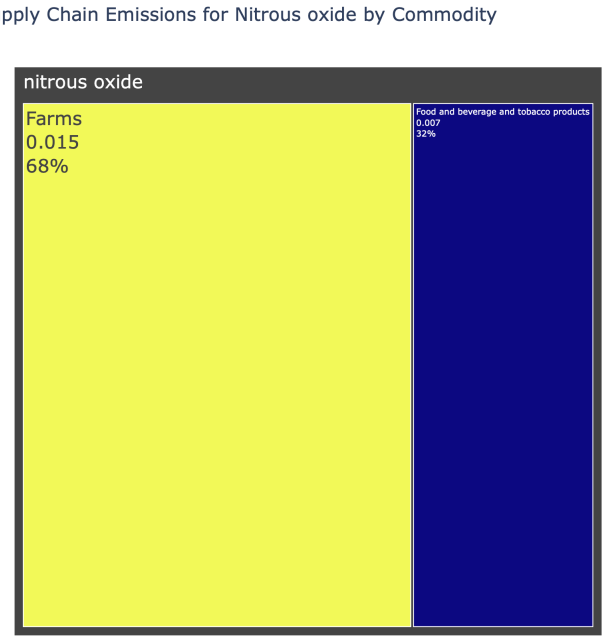


Fig. 19. Nitrous Oxide emissions across different commodities

- **Farms (68%):** Confirmed as the largest contributor, reinforcing the need to focus on emission reduction strategies in agriculture.
- **Food, Beverage, and Tobacco Products (32%):** Though smaller in magnitude, this category still represents a significant share of emissions, emphasizing the need for improvements in industrial processes.

The treemap refines the understanding by presenting precise emission values and percentages, reinforcing the dominance of farms in N_2O emissions.

Revised Narrative

The combined analysis underscores that agricultural activities are the primary drivers of N_2O emissions, making up more than two-thirds of the total. The contribution of food and beverage industries is notable but secondary, suggesting targeted strategies for emission reductions in both areas.

Other GHG Emissions—Refined Analysis

Refined Insights from the Treemap

The treemap (Fig. 20) consolidates these categories into broader sectors, providing a clearer view of contributions:

- **Transportation (60%):** Combining water and truck transportation, this category emerges as the dominant source of emissions, underscoring the environmental impact of the logistics industry.
- **Machinery (22%):** Reflecting emissions from manufacturing equipment and industrial processes, this remains a significant contributor.
- **Primary Metals (18%):** Highlighting the carbon-intensive nature of metal extraction and production.

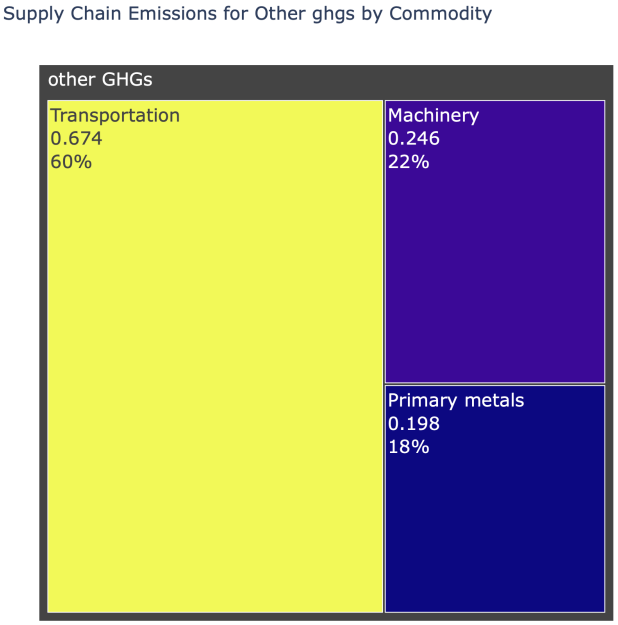


Fig. 20. Other GHG emissions across different commodities

The treemap not only consolidates the emissions sources but also provides precise values, reinforcing the outsized role of transportation.

Revised Narrative

The transition from the sunburst chart to the treemap shifts the focus from individual sectors to consolidated categories, revealing transportation's overwhelming contribution to other GHG emissions. The significant roles of machinery and primary metals further illustrate the broad industrial impact on emissions.

Iteration 1 Subflow 1 (Trend over time)

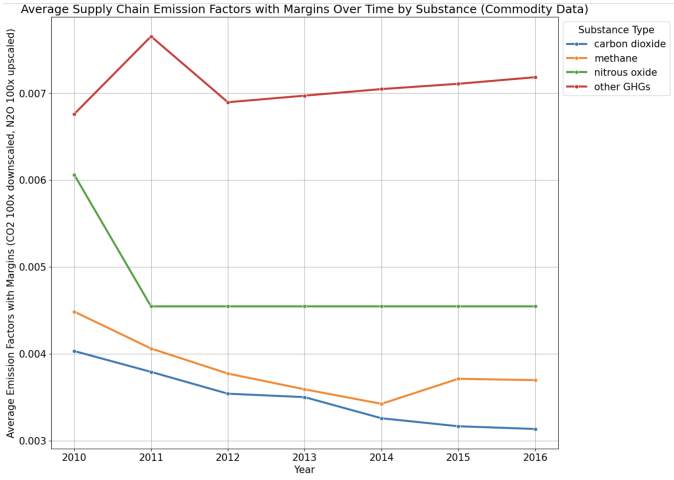


Fig. 21. Trend of average supply chain emission factors

Introduction to the Visualization

The visualization (Fig. 21) examines the average supply chain emission factors for various greenhouse gases (GHGs) from 2010 to 2016. These factors represent the environmental impact of commodities, expressed in terms of emissions of substances like carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and other GHGs. Margins are applied to normalize and highlight variations between the substances.

Key Observations

- **Carbon Dioxide (CO_2):** Displays a steady decline over the years, but a plateau in 2013.
- **Methane (CH_4):** Initially shows a significant decrease until 2014, followed by a jump in emissions for year 2015.
- **Nitrous Oxide (N_2O):** Experiences an initial sharp drop in year 2011, after which it remains consistent, hinting at stabilization after early reductions.
- **Other GHGs:** Stand out with a unique sharp inrease in year 2011 and an upward trend after 2012, diverging from the decline seen in other substances, highlighting an area needing focused intervention.

Iteration 2 Subflow 1 (Trend over time)

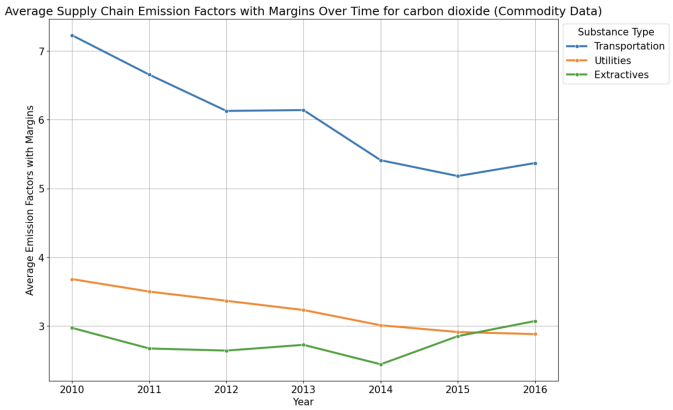


Fig. 22. Trend of supply chain emission factors for carbon dioxide

Investigating the 2013 Plateau in CO_2 Emissions

Introduction to the Visualization

This visualization (Fig. 22) focuses specifically on the top three contributors to carbon dioxide (CO_2) emissions: Transportation, Utilities, and Extractives, over the years 2010 to 2016. The purpose is to delve deeper into the observed plateau in CO_2 emissions between 2012 and 2013.

Explanation of the 2013 Plateau

The increase in Extractives emissions during 2013, coupled with the stability in Transportation emissions, appears to counterbalance the decline in emissions from Utilities. This equilibrium across the three sectors explains the "no change" observed in overall CO_2 emissions from 2012 to 2013.

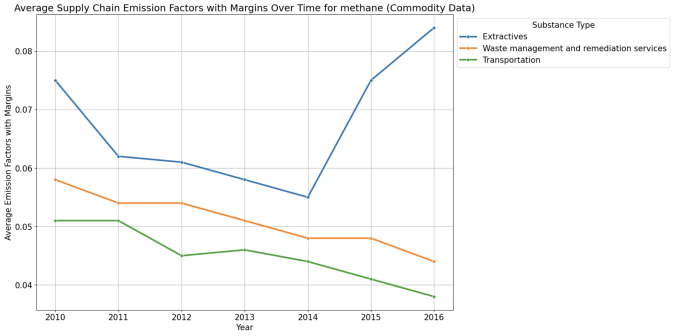


Fig. 23. Trend of supply chain emission factors for methane

Abrupt Increase in Methane Emissions in 2015

Introduction to the Dataset

This visualization (Fig. 23) focuses on methane (CH_4) emissions from the top three contributing sectors: Extractives, Waste Management and Remediation Services, and Transportation, over the years 2010 to 2016. The objective is to explain the sharp increase in methane emissions observed in 2015.

Explanation of the 2015 Spike

The extraordinary increase in methane emissions from Extractives in 2015 offsets the steady reductions from Waste Management and Transportation. The surge in Extractives emissions is so significant that it not only reverses the trend of declining methane levels but also results in a net increase, causing overall emissions to exceed prior years.

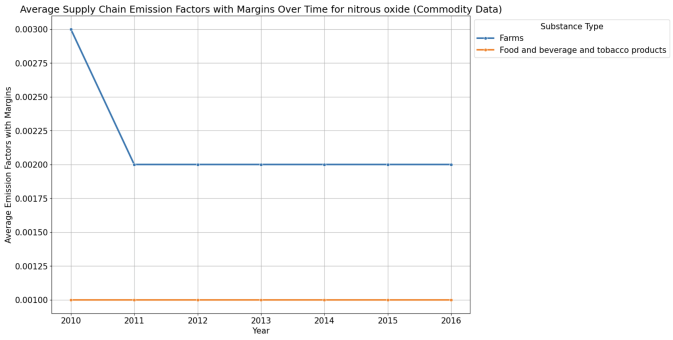


Fig. 24. Trend of supply chain emission factors for nitrous oxide

Sharp Decline in Nitrous Oxide Emissions in 2011

Introduction to the Dataset

This visualization (Fig. 24) focuses on nitrous oxide (N_2O) emissions from the top two contributors: Farms and Food, Beverage, and Tobacco Products, from 2010 to 2016. The objective is to explain the sharp decline in N_2O emissions observed in 2011 and its subsequent stabilization.

Explanation of the 2011 Decline

The sharp decline in N_2O emissions in 2011 is almost entirely driven by Farms, likely due to:

- Changes in agricultural practices, such as reduced fertilizer usage or the adoption of emission-reducing technologies.
- Policy interventions targeting the agriculture sector to curb nitrous oxide emissions.

Post-2011 stability suggests that these changes became permanent or routine, preventing further drastic reductions.

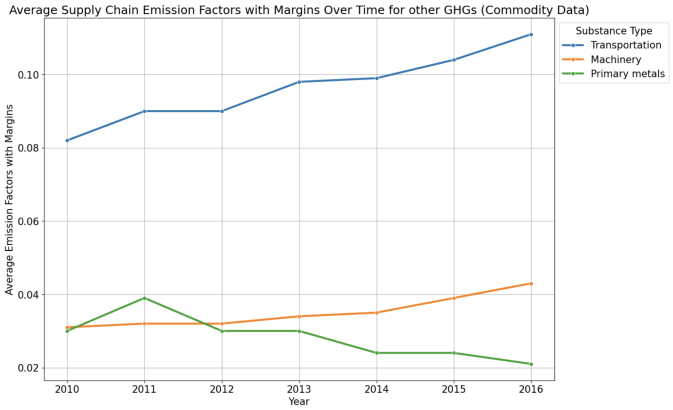


Fig. 25. Trend of supply chain emission factors for other GHGs

Abrupt Increase in Other GHG Emissions in 2011

Introduction to the Dataset

This visualization (Fig. 25) focuses on the emissions of other greenhouse gases (GHGs) from the top three contributing sectors: Transportation, Machinery, and Primary Metals, over the years 2010 to 2016. The objective is to explain the abrupt increase in other GHG emissions observed in 2011.

Explanation of the 2011 Spike

The abrupt increase in GHG emissions in 2011 is largely attributed to the sudden surge in emissions from Primary Metals.

- The consistent upward trends in Transportation and Machinery emissions further exacerbate the situation, leading to a cumulative impact.
- The combination of these factors results in a noticeable spike in 2011, marking a deviation from the otherwise gradual increase.

III. Task 3: Deeper Analysis into Specific Sectors' Emissions

This section builds upon the Supply Chain emission factors' plots analyzed in Assignment-1. The visualizations from Assignment-1 provided an overview of emission trends across various sectors, highlighting key contributors and patterns. However, to gain deeper insights and refine these analyses, this section incorporates additional datasets, leveraging *SubFlow-2* of the workflow. By integrating supplementary data, we aim to uncover sector-specific details and refine the understanding of emission behaviors.

One of the focal points of this section is applying clustering techniques to analyze vehicle emission data from the

additional datasets. Clustering enables grouping of emission patterns across different vehicle types or categories, revealing trends or anomalies that might not be evident from aggregate data. By combining the enriched datasets with clustering and iterative feedback, this section seeks to generate actionable insights into sectoral emission variability and key drivers. This iterative exploration enhances the comprehensiveness of the analysis, facilitating targeted recommendations for emission reduction strategies.

Utilities Splitted Emissions Analysis

The plot above illustrates the yearly emissions from the utilities sector, split by different greenhouse gases (GHGs) including Carbon Dioxide (CO_2), Methane (CH_4), Nitrous Oxide (N_2O), and Other GHGs from 2010 to 2016. The emissions are measured in total emissions per kilogram (kg) per 2018 USD purchaser price. Each GHG is represented by a distinct color, enabling a clear view of their relative contributions over time.

Fig. 26. Year-wise emission contributions in the Utilities sector

Plot Explanation: - The largest contributor to emissions is Carbon Dioxide (CO_2), which dominates the total emissions in each year, albeit its values are divided by 100 for visual clarity. - Methane and Nitrous Oxide emissions are relatively smaller but remain consistent throughout the period. - Other GHGs contribute to the emissions but show a stable or slightly declining trend over the years. - The total emissions exhibit a slight decrease from 2010 to 2016, indicating potential improvements in efficiency or emission reduction practices within the utilities sector.

Inferences: 1. CO_2 is the primary driver of emissions in the utilities sector, suggesting that strategies to reduce overall emissions should prioritize reducing CO_2 output. 2. The stable trends of Methane and Nitrous Oxide imply that while they are less significant contributors, they should not be ignored in mitigation efforts. 3. The slight decline in total emissions may indicate early signs of progress toward emission reduction goals.

Workflow Context: This plot was created using *SubFlow-1* of the workflow, which focuses on refining visualizations to highlight key patterns and trends. While SubFlow-1 provided an overview of emissions and their split across GHGs, this section will now leverage *SubFlow-2* to integrate additional datasets for deeper insights. By incorporating supplementary data, we aim to explore underlying factors driving these trends and identify opportunities for targeted interventions in the utilities sector.

Reference and Further Exploration

According to this article available on the internet - UTILITIES, the utilities sector encompasses the stock of companies operating in electric, gas, and water utilities. This definition provides a clear scope for analyzing emissions associated with these essential services. Building upon this understanding, we can leverage the additional dataset available through *SubFlow-2*, which contains emission factors for CO_2, CH_4, and other gases categorized by electricity boards.

Given the information from the bar chart above, which highlights CO_2 as the most significant contributor to emissions in the utilities sector, this analysis will focus solely on the CO_2 data. Using this data, I will create a choropleth map to visualize the regional distribution of CO_2 emission factors across electricity boards. This approach will provide insights into geographical variations in emissions, enabling targeted recommendations for emission reductions at the board level.

Iteration-2: Using SubFlow-2: To gain deeper insights into the utilities sector emissions, this iteration leverages additional datasets available through *SubFlow-2*. The dataset shown above contains emission factors categorized by electricity boards (subregions) and specifies the contributions of different greenhouse gases, namely Carbon Dioxide (CO_2), Methane (CH_4), and Nitrous Oxide (N_2O).

Dataset Description: The dataset includes the following key components:

- **Subregion:** Represents specific electricity boards or operational areas (e.g., AKGD for ASCC Alaska Grid, CAMX for WECC California). Each subregion corresponds to a specific geographical area of operation.
- **CO_2 Factor (lb CO_2/MWh):** Indicates the emission factor for Carbon Dioxide per megawatt-hour of electricity. This is the most significant emission factor in the dataset.
- **CH_4 Factor (lb CH_4/MWh):** Represents the emission factor for Methane per megawatt-hour of electricity. Methane, while less in quantity, has a higher global warming potential than CO_2.
- **N_2O Factor (lb N_2O/MWh):** Provides the emission factor for Nitrous Oxide per megawatt-hour of electricity. Although a minor contributor by weight, it has significant climate impacts.

The dataset comprehensively covers subregions across the United States, offering a granular view of emission factors by operational area. This information provides a strong foundation for visualizing regional variations and identifying high-emission subregions.

Choropleth Map: CO_2 Emissions by Subregion (2011): The above choropleth map visualizes the geographical distribution of CO_2 emission factors (in lb CO_2/MWh) across U.S. subregions for the year 2011. Each state is color-coded according to the CO_2 emission factor, with darker shades representing higher emission factors.

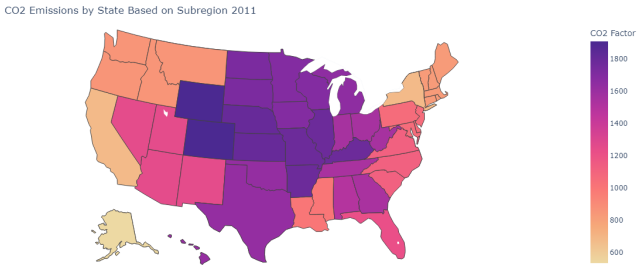


Fig. 27. Chloropleth of region wise C02 emission values in the US in 2024

Key Observations:

- Subregions like *RMPA (WECC Rockies)* and *SPNO (SPP North)* exhibit the highest emission factors, as seen by their darker shades.
- Coastal regions, particularly in the Northeast (*NPCC New England*) and West (*WECC Northwest*), have comparatively lower emission factors.
- Central regions display a mix of medium to high emission factors, highlighting variability across subregions.

This map provides valuable insights into regional variations in CO_2 emissions, enabling the identification of high-emission subregions for targeted mitigation strategies. The data-driven visualization emphasizes the need for region-specific policies to reduce CO_2 emissions.

Choropleth Map: CO_2 Emissions by Subregion (2024): The above choropleth map represents the CO_2 emission factors (in lb CO_2/MWh) across U.S. subregions for the year 2024. Compared to the 2011 choropleth, several changes in regional emission factors are evident.

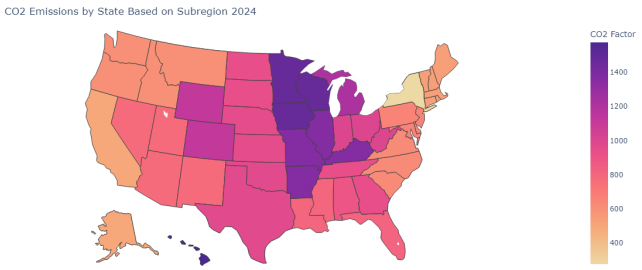


Fig. 28. Chloropleth of region wise C02 emission values in the US in 2024

Comparative Insights:

- Subregions like *RMPA (WECC Rockies)* and *RFCM (RFC Michigan)* continue to exhibit high CO_2 emission factors, maintaining their position as major contributors.
- Coastal regions (*NPCC New England* and *WECC Northwest*) retain relatively lower emission factors, indicating consistent emission control efforts in these areas.
- A notable reduction in emission factors is observed in subregions like *NYUP (NPCC Upstate NY)* and *SPNO (SPP North)*, suggesting successful implementation of emission mitigation strategies in these areas.
- Some central regions exhibit an increase in emission factors compared to 2011, emphasizing the need for renewed focus on emission reduction in these subregions.
- **Overall Insight:** There has been a general decline in CO_2 emission factors across the electricity sector in the U.S. supply chain, reflecting progress in reducing emissions at a national level.

This comparative analysis highlights regional trends over time, enabling the identification of areas with successful mitigation efforts and those requiring additional attention. The overall decline in CO_2 emissions in the electricity sector indicates positive advancements, although further work is required to sustain and amplify these efforts.

Verification with External Data: The findings of this analysis, which indicate a general decline in CO_2 emissions in the electricity sector between 2011 and 2024, are consistent with an article published by the U.S. Energy Information Administration (EIA) - LINK. The article provides an overview of total CO_2 emissions across various sectors, including transportation, electric power, industrial, residential, and commercial, as shown in the figure below.

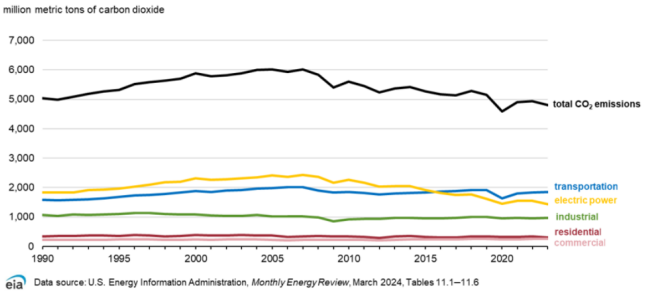


Fig. 29. Trends in CO_2 emissions by sector from 1990 to 2024. Source: U.S. Energy Information Administration, Monthly Energy Review, March 2024.

Consistency with Our Findings: As observed in Figure 29, CO_2 emissions in the electric power sector (yellow line) have shown a consistent decline over the years, particularly between 2011 and 2024. This aligns with our analysis of subregional data, which revealed reductions in CO_2 emission factors for several regions over the same period. Furthermore, the total CO_2 emissions (black line) have also decreased, reinforcing the trend of improving emission controls across sectors.

This external validation strengthens the reliability of our findings, highlighting that the reduction in supply chain emissions from the electricity sector is part of a broader national trend. Such insights underline the progress made in emission mitigation strategies and the importance of sustained efforts to achieve further reductions.

Heatmap of Emissions by Industry and Gas Type

This section follows *SubFlow-2*, similar to the previous part, to gain deeper insights into emissions data. The heatmap be-

low, was created in Assignment-1 and it provides an overview of emissions from different gases across various industries, highlighting their relative contributions.

Fig. 30. Heatmap showing emissions of different gases across industries. Emission values are scaled for comparability, with Carbon Dioxide (CO_2) as the primary contributor.

Explanation of the Heatmap

The heatmap visualizes the emission levels of four greenhouse gases—Carbon Dioxide (CO_2), Methane (CH_4), Nitrous Oxide (N_2O), and Other GHGs—across several industries. The intensity of the colors represents the magnitude of emissions, with brighter colors indicating higher values. The values are scaled for comparability:

- **Carbon Dioxide (CO_2):** The most significant contributor, with notable values in industries such as *Air Transportation* (0.872) and *Waste Management* (0.565).
- **Methane (CH_4):** Observed predominantly in *Farms* (0.51) and *Waste Management* (0.29).
- **Nitrous Oxide (N_2O):** Emissions are negligible across most industries, reflecting limited impact.
- **Other GHGs:** Moderate emissions are seen in *Primary Metals* (0.21) and *Farms* (0.15).

Emphasis on Air Transportation

One of the most striking observations from the heatmap is the exceptionally high CO_2 emission value in the *Air Transportation* industry (0.872). This value significantly surpasses other industries, indicating a critical area of focus for emission reduction strategies. Methane and Other GHGs show minimal contributions for this industry, reinforcing the dominance of CO_2 in air transportation emissions.

While this heatmap provides a high-level view of industry-specific emissions, we will now leverage *SubFlow-2* to enrich this analysis. By incorporating additional datasets, we aim to create new visualizations to explore emissions data further, particularly focusing on the high CO_2 emissions in the *Air Transportation* industry. This enhanced analysis will enable a deeper understanding of the drivers behind these emissions and inform strategies for targeted interventions.

Dataset Description

The dataset is one of the tables in the 3 datasets which are additional to the originally used ones and it provides detailed information on CH_4 (methane) and N_2O (nitrous oxide) emission factors for various vehicle types and years, specifically for mobile combustion sources. Below is a detailed description of the dataset:

Key Attributes:

- **Vehicle Type:** Describes the type of vehicle or equipment (e.g., LPG Non-Highway Vehicles, Diesel Ships and Boats, Jet Fuel Aircraft).
- **CH_4 Factor (g/gallon):** Indicates the methane emission factor, measured in grams per gallon of fuel consumed.
- **N_2O Factor (g/gallon):** Indicates the nitrous oxide emission factor, measured in grams per gallon of fuel consumed.
- **Year:** Specifies the year (2011 or 2014) for which the emission factors were measured.

Iteration-2 SubFlow-2: Scatter Plot Matrix (SPLOM)

This iteration follows *SubFlow-2* to explore the relationships between different emission factors across vehicle types using a Scatter Plot Matrix (SPLOM). The plot showcases pairwise comparisons of CH_4 and N_2O factors across all vehicle types. Each diagonal element displays a Kernel Density Estimate (KDE) plot for the corresponding factor, illustrating its distribution, while the off-diagonal elements present scatter plots to show relationships between these factors.

Explanation of the Plot: - The KDE plots on the diagonal highlight the distribution of CH_4 and N_2O factors. CH_4 has a highly skewed distribution due to the presence of an outlier with an exceptionally high value. - The scatter plots on the off-diagonal indicate a weak relationship between CH_4 and N_2O emissions, as most points are scattered without a clear trend.

Outlier Analysis: One striking observation from the SPLOM is the presence of an extreme outlier in the CH_4 factor, corresponding to the *Aviation Gasoline Aircraft* vehicle type. This vehicle type exhibits an exceptionally high CH_4 emission factor (approximately 7.04 g/gallon), significantly exceeding the values for other vehicle types. This outlier's presence skews the overall distribution of CH_4 emissions, as evident in the KDE plot, and underscores its critical impact on overall emission trends.

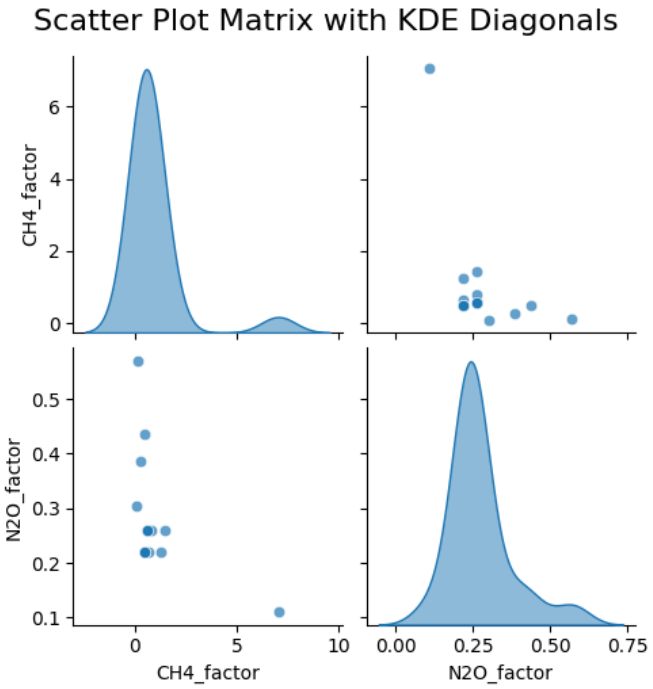


Fig. 31. SPLOM showing CH_4 and N_2O factors of different Vehicle Types

Insights and Next Steps: The SPLOM plot effectively visualizes the variability in CH_4 and N_2O emissions across vehicle types, allowing for a deeper exploration of their relationships. The identification of the *Aviation Gasoline Aircraft* as an outlier highlights the need for targeted interventions to address its disproportionately high emissions. Moving forward, we can leverage *SubFlow-2* to incorporate additional datasets and refine visualizations, enabling further insights into the underlying factors contributing to these emission patterns.

Iteration 3: SubFlow-1: Parallel Coordinates Plot (PCP)

This iteration follows *SubFlow-1*, where we use the additional dataset to create a new visualization, in this case, a Parallel Coordinates Plot (PCP). SubFlow-1 emphasizes the use of the original or additional datasets to derive insights through visual analytics. This approach allows us to explore data relationships and trends interactively.

Explanation of the Plot: The PCP visualizes the average of CH_4 and N_2O factors for each vehicle type across the years 2011, 2014, and 2015. Each axis represents the combined average of CH_4 and N_2O for a specific year, while the lines connecting the axes represent the trends for individual vehicle types.

Key observations from the plot include:

- A clear outlier is present, corresponding to a vehicle type with exceptionally high average values, which dominates the plot.
- For most vehicle types, the average values of CH_4 and N_2O remain clustered near the lower end of the axis, making it challenging to distinguish individual trends.

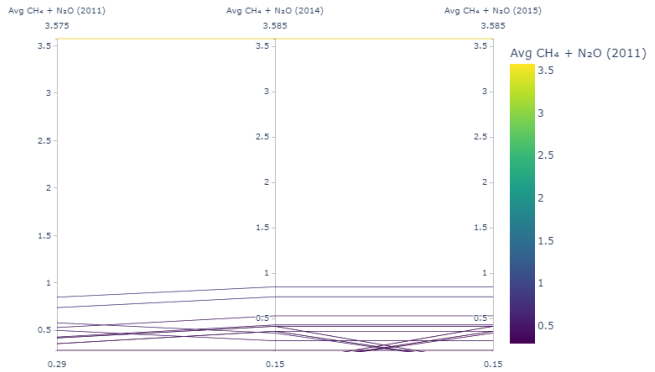


Fig. 32. Parallel Coordinates Plot: Average CH_4 and N_2O Factors Across Years

Challenges in Interpretation: The presence of an outlier skews the y-axis, resulting in a crammed display of other data points. This reduces the interpretability of the plot, as the visualization fails to provide clear insights into the trends for vehicle types with lower average values.

Next Steps: To address this issue, the next iteration will improve the visualization by applying a logarithmic transformation to the average values, thereby reducing the impact of the outlier and enhancing the overall readability of the plot.

Iteration 4: SubFlow-1: Improved Parallel Coordinates Plot with Log Transformation

This plot represents an enhanced version of the Parallel Coordinates Plot, achieved through the application of a log transformation to the CH_4 and N_2O emission factors. The log transformation addresses key issues from the previous iteration, resulting in a more interpretable and visually streamlined representation of the data.

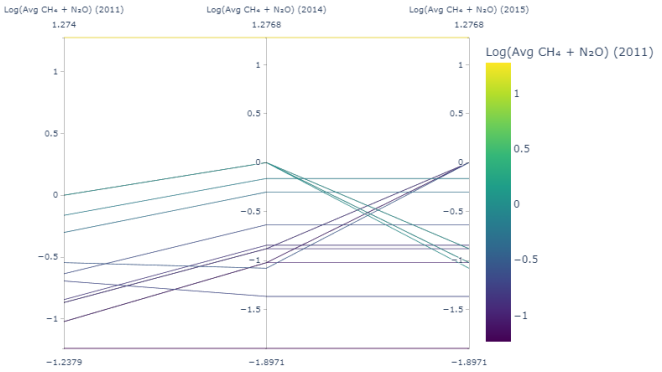


Fig. 33. Parallel Coordinates Plot: Log transformed Average CH_4 and N_2O Factors Across Years

Key Improvements:

- **Log Transformation:** By applying a log transformation to the emission factors, the plot effectively handles outliers, compressing their impact and ensuring that extreme values do not dominate the visualization.

- **Reduction of Visual Clutter:** The previous plot suffered from overlapping lines and an inability to clearly differentiate data trends. This improved plot mitigates such issues, making it easier to identify patterns and relationships.
- **Interpretable Axes:** While the y-axis values are negative, this does not imply negative emissions. Instead, these values reflect emissions below 1 in the original scale, as they are log-transformed. This is crucial for understanding emission trends in a more standardized way.

Interpretation:

- The plot highlights how CH_4 and N_2O emissions vary across different years, with a smooth progression of values.
- The use of color encoding for 2011 log-transformed values provides an additional layer of insight, allowing for quick identification of higher and lower emissions across different vehicle types.

Conclusion: This improved Parallel Coordinates Plot demonstrates the value of log transformation in emission analysis. By handling outliers and reducing visual clutter, it enables a more effective and accurate exploration of emission trends over time. This iteration underscores the importance of thoughtful data transformation techniques in complex visualizations.

Validation of Inference Consistency Using Historical Passenger Traffic Data

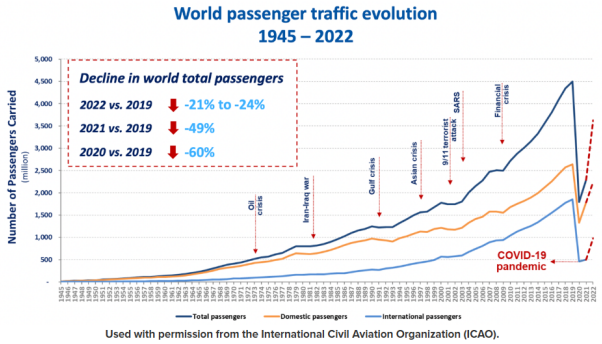


Fig. 34. World Passenger Traffic Evolution (1945–2022). Source: The Growth in Greenhouse Gas Emissions from Commercial Aviation (2019, updated 2022).

Key Points from the Plot:

- The graph shows the evolution of world passenger traffic (in millions) from 1945 to 2022, segmented into total passengers, domestic passengers, and international passengers. Source - ARTICLE
- Key global crises are marked, including the oil crisis, Iran-Iraq war, 9/11 terrorist attacks, SARS, financial crisis, and most recently, the COVID-19 pandemic.
- The COVID-19 pandemic caused a sharp decline in passenger numbers, with reductions of:
  - 60% in 2020 compared to 2019,
  - 49% in 2021 compared to 2019,
  - 21–24% in 2022 compared to 2019.

This plot substantiates the consistency and validity of our analysis. It demonstrates how external factors like crises and pandemics impact aviation passenger trends, which in turn directly influence the trajectory of aviation-related emissions.

Scatter Plot Matrix for CH_4 and N_2O Emissions and Clustering Techniques

Introduction:

In this final section of our deep-dive analysis, under SubFlow-1, we shall create a scatter plot matrix to analyze the relationship between CH_4 (methane) and N_2O (nitrous oxide) emissions. The focus is on 'Mobile Combustion CH_4 and N_2O for On-Road Gasoline Vehicles.' This visualization allows us to explore potential correlations and identify patterns or outliers that could provide insights into emission behaviors across vehicle types.

Dataset Description

This dataset provides detailed information about CH_4 (methane) and N_2O (nitrous oxide) emission factors for various types of gasoline-powered vehicles. The emission factors are expressed in grams per vehicle-mile and are categorized by vehicle type. Below is a summary of the dataset:

Key Attributes:

- **Vehicle Type:** Specifies the category of the vehicle, such as Passenger Cars, Light-Duty Trucks (e.g., vans, pickup trucks, SUVs), Heavy-Duty Vehicles, and Motorcycles.
- **CH_4 Factor (g CH_4 / vehicle-mile):** Represents the methane emission factor, measured in grams per vehicle-mile. This factor varies by vehicle type and operational characteristics.
- **N_2O Factor (g N_2O / vehicle-mile):** Represents the nitrous oxide emission factor, measured in grams per vehicle-mile. Like methane, this factor is dependent on the vehicle type and other influencing factors.

Scatter Plot Matrix of CH_4 and N_2O Emission Factors

The scatter plot matrix (SPLOM) provides a detailed visualization of the pairwise relationships between CH_4 (methane) and N_2O (nitrous oxide) emission factors. This visualization is key for identifying patterns, distributions, and potential clusters in the dataset.

Key Observations:

- **Diagonal KDE Plots:**
  - The **CH_4 factor** demonstrates a right-skewed distribution, with most values concentrated below 0.2.
  - The **N_2O factor** shows a steep peak around 0.05, tapering off towards 0.15.
- **Scatter Plots:**
  - The off-diagonal scatter plots indicate a nonlinear relationship between CH_4 and N_2O emissions, with points clustering along curved trends.

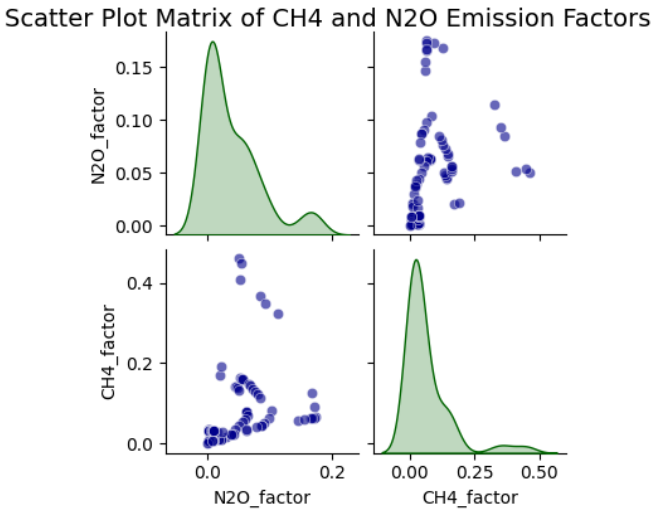


Fig. 35. Scatter Plot Matrix of CH_4 and N_2O Emissions for On-Road Gasoline Vehicles.

– Data points are concentrated in specific regions of the plot, suggesting distinct emission profiles for certain vehicle types or operational conditions.
– The points are visually enhanced with a consistent `darkblue` color and partial transparency (`alpha=0.6`) to improve readability and distinguish overlapping data.

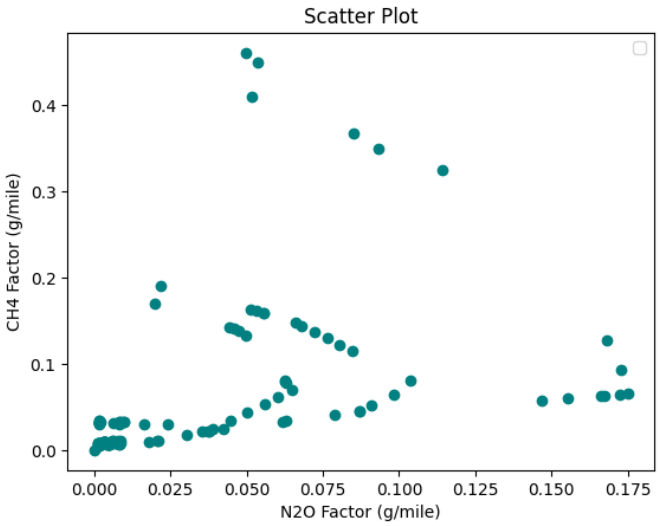


Fig. 36. Scatter Plot of CH_4 and N_2O Emissions for On-Road Gasoline Vehicles.

Improvements in This Plot:

- Enhanced KDE plots with smoother curves, clearly highlighting the density distribution of emission factors.
- A more balanced layout for better clarity and separation of data trends.

Conclusion: The scatter plot matrix serves as an exploratory tool to understand the emission patterns of CH_4 and N_2O. The presence of nonlinear relationships and distinct groupings in the data provides a foundation for further analysis. As a logical next step, we propose employing a clustering algorithm to categorize the points and uncover deeper insights into the emission behaviors of different vehicle types.

Elbow Method to Determine Optimal Clusters

The elbow method is a commonly used technique to determine the optimal number of clusters for a dataset in clustering algorithms, such as K-means. It works by plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the point where the rate of decrease slows down, forming an ”elbow.” This point indicates the optimal number of clusters that balance complexity and performance.

We will apply the elbow method to our dataset to identify the optimal number of clusters for grouping vehicles based on their CH_4 and N_2O emission factors. This will enable us to train a model that effectively categorizes vehicles into meaningful clusters, aiding in further analysis and insights.

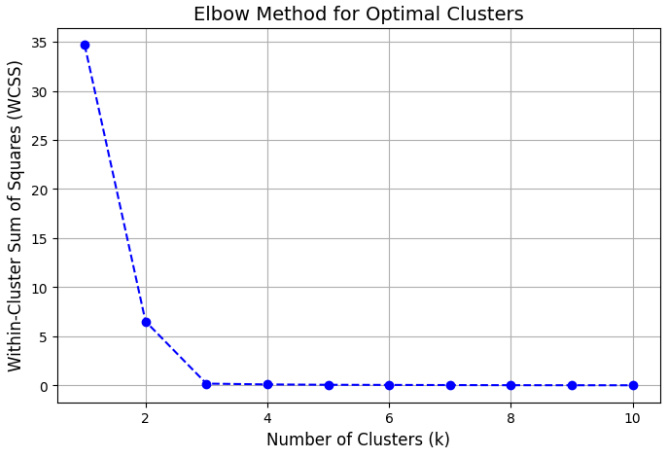


Fig. 37. Elbow method graph of Scatter Plot between CH_4 and N_2O Emissions for On-Road Gasoline Vehicles.

The elbow method graph illustrates the relationship between the number of clusters and the Within-Cluster Sum of Squares (WCSS). From the graph, the following observations can be made:

- There is a sharp decrease in WCSS as increases from 1 to 3, indicating that adding more clusters significantly improves clustering performance during this range.
- Beyond , the rate of decrease in WCSS slows down considerably, forming a distinct ”elbow” at .

Based on the elbow method graph, the optimal number of clusters for this dataset is 3. This choice ensures a balance between minimizing within-cluster variance and avoiding unnecessary complexity in the clustering model. The next step will involve training the clustering model with to group the vehicles effectively based on their emission factors.

K-Means Clustering on Scatter Plot

Clustering is a powerful technique used to group similar data points based on their characteristics. By applying K-

Means clustering to this dataset, we aim to identify natural groupings of vehicles based on their CH_4 (methane) and N_2O (nitrous oxide) emission factors. These clusters can provide valuable insights into patterns and relationships within the data, potentially revealing distinct emission profiles for different types of vehicles.

Objective: The primary objective of applying K-Means clustering is to:

- Group vehicles into clusters with similar emission characteristics.
- Understand how CH_4 and N_2O emissions vary across different groups.
- Provide actionable insights for targeting specific vehicle groups in emission reduction strategies.

Approach: Using the optimal number of clusters ($k = 3$), determined through the elbow method, we apply K-Means clustering on the scatter plot of CH_4 and N_2O emission factors. Each data point in the scatter plot is assigned to one of the three clusters, based on its proximity to the cluster centroids.

The resulting clusters can be visualized on a scatter plot with distinct colors for each cluster, allowing for a clear and intuitive interpretation of the grouping.

Next Steps: The clustered scatter plot will be analyzed to identify key trends and characteristics within each cluster. This analysis will help in formulating hypotheses about the underlying factors contributing to emissions and inform strategies for reducing vehicle emissions effectively.

K-Means Clustering on Scatter Plot

Clustering is a powerful technique used to group similar data points based on their characteristics. By applying K-Means clustering to this dataset, we aim to identify natural groupings of vehicles based on their CH_4 (methane) and N_2O (nitrous oxide) emission factors. These clusters can provide valuable insights into patterns and relationships within the data, potentially revealing distinct emission profiles for different types of vehicles.

Objective: The primary objective of applying K-Means clustering is to:

- Group vehicles into clusters with similar emission characteristics.
- Understand how CH_4 and N_2O emissions vary across different groups.
- Provide actionable insights for targeting specific vehicle groups in emission reduction strategies.

Approach: Using the optimal number of clusters ($k = 3$), determined through the elbow method, we apply K-Means clustering on the scatter plot of CH_4 and N_2O emission factors. Each data point in the scatter plot is assigned to one of the three clusters, based on its proximity to the cluster centroids.

The resulting clusters can be visualized on a scatter plot with distinct colors for each cluster, allowing for a clear and intuitive interpretation of the grouping.

K-Means Clustering Output and Cluster Analysis

The K-Means clustering algorithm was applied to the dataset using the optimal number of clusters ($k = 3$) determined by the elbow method. The resulting scatter plot, shown above, visualizes the CH_4 (methane) and N_2O (nitrous oxide) emission factors for different vehicles, grouped into three distinct clusters. Each cluster is represented by a unique color, with red crosses marking the cluster centroids.

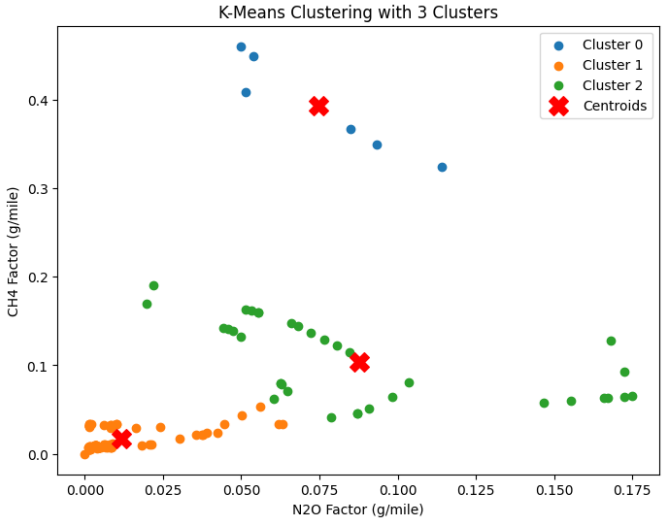


Fig. 38. K-Means Clustering of Vehicles Based on CH_4 and N_2O Emission Factors. The scatter plot groups vehicles into three clusters: Cluster 0 (Old Cars), Cluster 1 (Middle-generation Cars), and Cluster 2 (Latest Cars), with red crosses indicating the cluster centroids. Each cluster represents distinct emission profiles, highlighting the progression in vehicle emission control technologies.

Cluster Names and Interpretation: To better interpret the clusters, they are categorized as follows:

- **Cluster 0 (Old Cars):** Vehicles in this cluster generally exhibit **high emission factors** for both CH_4 and N_2O. These vehicles likely represent older models with outdated emission control technologies.
- **Cluster 1 (Middle-generation Cars):** Vehicles in this cluster have **moderate CH_4** emissions and **relatively low N_2O** emissions. They likely represent mid-range models with improved, yet not fully modern, emission controls.
- **Cluster 2 (Latest Cars):** This cluster contains vehicles with **low emissions** for both CH_4 and N_2O, representing the latest models equipped with advanced emission control systems.

Cluster Characteristics:

- **Old Cars (Cluster 0):**
  - High CH_4 emission factors, often exceeding 0.3 g/mile.
  - Elevated N_2O emission factors, clustering around 0.1 g/mile.
  - These vehicles are likely older and less efficient, contributing significantly to greenhouse gas emissions.
- **Middle-generation Cars (Cluster 1):**

  - Moderate CH_4 emission factors ranging between 0.05 and 0.15 g/mile.
  - Lower N_2O emissions compared to Cluster 0, typically below 0.05 g/mile.
  - Likely mid-generation vehicles with partial adoption of modern emission control technologies.
- **Latest Cars (Cluster 2):**
  - Very low CH_4 emission factors, typically below 0.05 g/mile.
  - Minimal N_2O emission factors, often below 0.02 g/mile.
  - These vehicles represent the most advanced and environmentally friendly models in the dataset.

Conclusion: The clustering output successfully groups vehicles into distinct categories based on their emission profiles. The clear distinction between the clusters highlights the technological evolution in vehicle emission control systems. This analysis provides a foundation for targeting high-emission vehicles (Cluster 0) for retrofitting or replacement, while emphasizing the success of modern emission standards seen in Cluster 2.

Validation of the Clustering Method Output

The findings from the clustering analysis, which categorized vehicles into three groups (Old Cars, Middle Cars, and Latest Cars), align with broader trends in vehicle emissions and advancements in technology. The graph in Figure 39, which is part of an article of a survey titled - 'Progress Cleaning the Air and Improving People's Health' (Source - ARTICLE) provides a clear validation of these findings. It shows the relationship between vehicle miles traveled and volatile organic compound (VOC) emissions over time, indicating a significant decline in emissions per mile despite an increase in miles traveled.

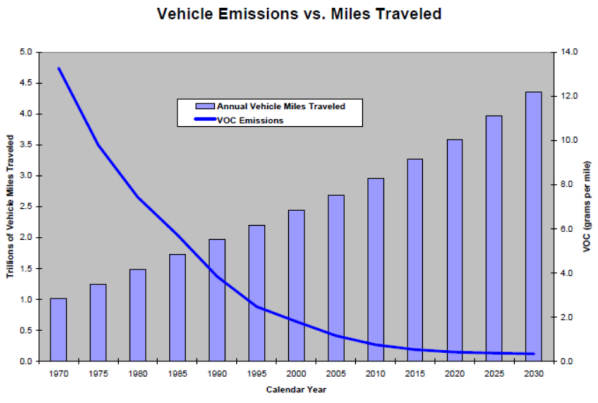


Fig. 39. Vehicle Emissions vs. Miles Traveled: The graph shows a steady decline in VOC emissions per mile despite an increase in total vehicle miles traveled, highlighting the impact of modern emission-reducing technologies.

Key Observations:

- From 1970 onwards, the VOC emissions (blue line) have steadily decreased, even as the annual vehicle miles traveled (blue bars) have increased. This trend reflects the adoption of modern technologies in vehicles, such as catalytic converters and stricter emission standards.
- The sharp decline in VOC emissions from 1970 to 2000 is consistent with the characteristics of vehicles in the "Old Cars" and "Middle Cars" clusters, where older vehicles contributed higher emissions, and newer mid-generation vehicles began incorporating emission-reducing technologies.
- After 2000, the VOC emissions plateau at a very low level, corresponding to vehicles in the "Latest Cars" cluster. These vehicles feature state-of-the-art emission control systems, leading to minimal environmental impact.

Conclusion: The clustering analysis is validated by this historical trend, which highlights the significant reduction in emission factors over time due to technological advancements. The clear separation of vehicle groups in the clustering analysis aligns with the progression of emission control technologies depicted in Figure 39. This consistency reinforces the reliability of the clustering method and its ability to reflect real-world changes in vehicle emissions.

Member-wise Contributions

- Task-1 : Krish Patel
- Task-2 : Vansh Sinha
- Task-3 : Ashirwad Mishra

References

1) Plotly Express Official Documentation, Scatter Plot Matrix. Available at: https://plotly.com/python/splom/
2) Plotly Express Official Documentation, Parallel Coordinates Plot. Available at: https://plotly.com/python/parallel-coordinates-plot/
3) Pandas Python Library Documentation. Available at: https://pandas.pydata.org/docs/
4) Seaborn Python Library Documentation, Pairplot and KDE Plots. Available at: https://seaborn.pydata.org/generated/seaborn.pairplot.html
5) GeoJSON and Choropleth Mapping Tutorial (Electricity Board Regions). Available at: https://plotly.com/python/choropleth-maps/
6) Matplotlib Official Documentation. Available at: https://matplotlib.org/stable/contents.html
7) U.S. Environmental Protection Agency, Progress Cleaning Air and Improving People's Health. Available at: ARTICLE
8) Environmental and Energy Study Institute, Growth in Greenhouse Gas Emissions from Commercial Aviation. Available at: ARTICLE
9) U.S. Energy Information Administration, U.S. Energy-Related CO2 Emissions Declined by 3% in 2023. Available at: LINK
10) Scikit-learn Documentation: K-means Clustering. Available at: https://scikit-learn.org/stable/modules/clustering.html#k-means

APPENDIX

The A-1 report has been added as appendix to this A-3 report. The images of the plots used in A-1 can be referenced from this appendix.

DAS732: Data Visualisation Assignment 1 Report

Ashirwad Mishra
IMT2022108
IIIT-Bangalore
Ashirwad.Mishra@iiitb.ac.in

Krish Patel
IMT2022097
IIIT-Bangalore
Krish.Patel@iiitb.ac.in

Vansh Sinha
IMT2022122
IIIT-Bangalore
Vansh.Sinha@iiitb.ac.in

I. Dataset Description

The Dataset used in this assignment is named "Supply Chain Greenhouse Gas Emission Factors for US Industries and Commodities" and is publicly available on this website -Dataset. This contains year-wise (2010-2016), Summary and Detailed versions of GHG emissions from Commodities and Industries in the US Economy. The fields/columns in the dataset are as follows:

1. Commodity Code - Code of the commodity or industry from the BEA Make and Use Tables 2012 categorization. 'Detail' and 'summary' are two levels of detail BEA publishes economic input-output accounts data at. 'Detail' level is the most resolved categorization and includes 405 commodity or industry sectors. 'Summary' level is a categorization with medium resolution and includes 73 commodity and 71 industry sectors.

2. Commodity Name - Name of the commodity or industry from the BEA Make and Use Tables 2012 categorization, except detail commodities, which use USEEIO v1.1 names, see Ingwersen and Yang 2017)

3. Substance - Greenhouse gas: 'carbon dioxide' is CO2; 'methane' is CH4; 'nitrous oxide' is N2O; and 'other GHGs' include HFC-23, HFC-32, HFC-125, HFC-134a, HFC-143a, HFC-236fa, CF4, C2F6, C3F8, C4F8, SF6, and NF3

4. Unit - Unit of emission factors for each gas. 'Other GHGs' are aggregated and reported in CO2e (carbon dioxide equivalents) using the IPCC AR4 100-year GWP factors. Purchaser price is the price paid by the consumer and equals to the producer prices plus any associated margin, which generally include distribution, wholesale and retail costs.

5. Supply Chain Emission Factors without Margins - Direct and indirect GHG emissions associated with production of commodity or industry from cradle to the point of production(kg) per 2018 USD of that commodity or industry in the US in purchaser price .

6. Margins of Supply Chain Emission Factors - Direct and indirect GHG emissions associated with production of commodity or industry from the point of production to the point of sale (kg) per 2018 USD of that commodity or industry in the US in purchaser price of that commodity or industry in the US.

7. Supply Chain Emission Factors with Margins - Direct and indirect GHG emissions associated with production of commodity or industry from cradle to the point of sale(kg) per 2018 USD of that commodity or industry in purchaser price of that commodity or industry in the US.

8. DQ ReliabilityScore of Factors without Margins - Data reliability scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

9. DQ TemporalCorrelation of Factors without Margins - Data temporal correlation scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

10. DQ GeographicalCorrelation of Factors without Margins - Data geographical correlatoin scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

11. DQ TechnologicalCorrelation of Factors without Margins - Data technological correlation scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

12. DQ DataCollection of Factors without Margins - Data collection scores for model results using USEPA 2016 Data quality assessment system, where 1 is the better quality and 5 the poorer quality

After cleaning the dataset and creating our dataframe for analysis, we created an additional column : Year - The year in which the data was recorded

II. Tasks

A. Task-1: Commodity-wise Emission Factors and Their Trends

Data Story: This section focuses on understanding the environmental impact of different commodities by analyzing their emission factors. It answers questions about which commodities contribute most to emissions and how these emissions have changed over time.

Sub-questions:

- What are the total emission factors (with margins) for each commodity?
- Which substances contribute the most to the emissions of each commodity (e.g., CO_2, methane)?
- How do emissions for different commodities evolve over time?
- Are there any seasonal or annual patterns in emissions for certain commodities?

B. Task-2: Substance-wise Analysis and Contribution to Global Emissions

Data Story: This section explores the dataset from a substance-based perspective, focusing on how different gases like CO2, methane, and nitrous oxide contribute to global emissions across various commodities.

Sub-questions:

- Which substances contribute the most to global emissions across all commodities?
- Which commodities/industries contribute most to emissions of a particular substance?
- What is the trend of specific substances (e.g., CO2) over the years across all commodities?

C. Task-3: Data Quality (DQ) Scores and Their Impact on Emission Estimates

Data Story: This section explores the data quality of emission estimates, focusing on how reliable the data is for each commodity and substance. It assesses the strength of the data and highlights areas where improvement is needed.

Sub-questions:

- Which commodities have the best and worst Data Quality (DQ) scores?
- What is the overall trend in data quality scores for different substances across all commodities?
- Is there a correlation between emission factors and DQ scores for different commodities?
- How do the various DQ score components (e.g., reliability, geographical correlation) differ amongst themselves?

III. Data Stories

In this section, we analyse in depth, the data stories created and try to answer the questions which arise from them.

A. Task-1 : Commodity-wise Emission Factors and Their Trends

We first try to plot the Total Supply Chain Emissions by commodities. We expect the Total Emissions to decrease over the time as the modern world centers itself around green energy. After the plot we do find that the data aligns with other prior assumption. Fig. 1 shows the Top 5 Commodities with highest Supply Chain Emissions (with margins). We can conclude that generally, most of the commodities have a decreasing trend in GHGs emissions. We can also see that "*Utilities*" commodity has significantly higher amounts of emissions compared to other commodities. Other inferences that we can make from the plot is the "*Petroleum and coal products*" commodity has seen a rise in Supply Chain Emissions since 2014.

We saw the data for commodities, let us now see what are the trends for the industry emissions over the last few years. Fig. 2 shows the Supply Chain Emissions by the top 5 highest contributors to Global emissions. We can see that the "*Utilities*" industry has higher emissions than "*Utilities*" commodity. The other trends are very similar to that of the commodity emission trends.

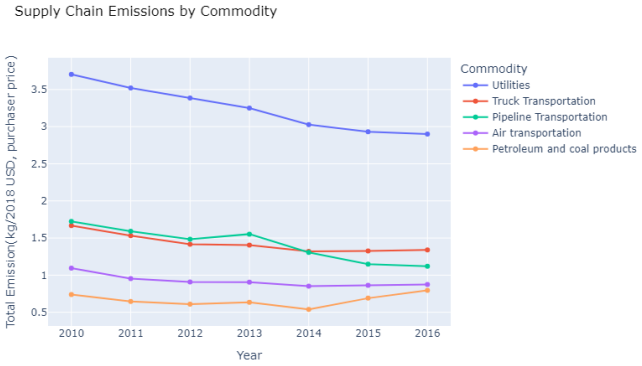


Fig. 1. Total Emissions for commodities

We have used line charts for Fig. 1 and 2 as we had to show data for many commodities and industries, and line chart also gives a better visualization of the decreasing trend in total emissions over the years. Markers are used to depict values for total emissions for every year on the x-axis.

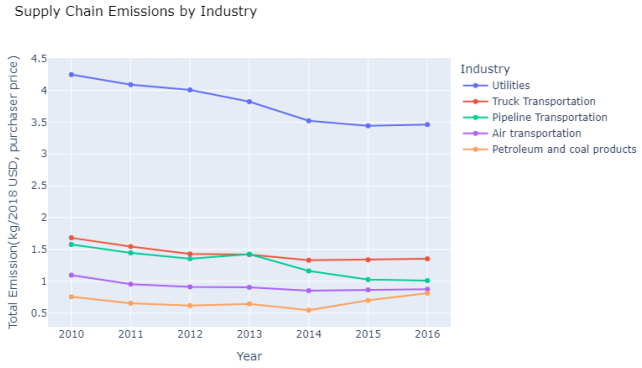


Fig. 2. Total Emissions for industries

We have already seen the trends for Supply Chain Emissions for commodities and industries, but it includes emissions of all Green House Gases. Let us now see how much each gas contributes to total emissions of an individual industry. Let us take the industry which has the highest amount of total emission, i.e, "*Utilities*".

Fig. 3 uses stacked Bar chart to show how much each gas contributes to the total emissions of the selected industry. You must note that the data for carbon dioxide gas is scaled down by 100 times. This information is strong enough to show that carbon dioxide is the most contributing gas to the total emissions. However, we also note that there is no emission of nitrous oxide by the selected industry. Here the stacked bar chart is used to as we had to show the split values of each gas and also their trends over the years.

Let us also pick a commodity and see how is the split of emissions for that particular commodity. Fig. 4 shows the split of emissions by gases for "Farms" commodity. Again, the

Fig. 3. Gas wise emissions of Utilities industry

values for carbon dioxide is scaled down by 100 to fit into the visualization. We can see that carbon dioxide and methane are the highest contributors for the selected commodity. For the particular visualization, we had a choice for either a stacked bar chart or a grouped bar chart but as the trends for individual gases were very much unpredictable, stacked bar chart could have given us wrong inferences and thus we choose grouped bar chart over stacked.

Now let us try to visualise the different gas emissions from different industries using Heat Maps. This would make it more easier for us to understand how much each gas contributes to total emissions of an industry. We take the most recent data of 2016 as it makes much more sense to see how the split for emissions is in the recent past. Fig. 5 is a Heat Map for the same. The values for methane, and other GHGs is scaled up by a factor of 10, while the values for nitrous oxide is scaled up by 100. This scaling factor itself speaks volumes about how tha split actually is in the recent past. We even see that most of the industries have 0 nitrous oxide emissions. Let me remind you all these values are in kg/2018 USD, Purchaser Price.

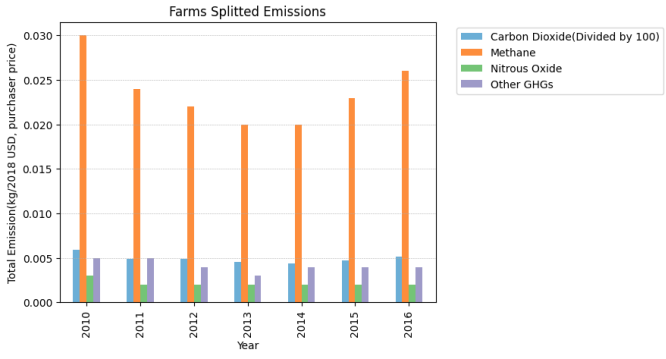


Fig. 4. Gas wise emissions for Farms commodity

A heat map is a perfect visualization for comparing multiple values against multiple attributes. In this case, as we had to show the emission data for different gases and different industries, we preferred a heat map. Heat map also gives a pretty much clear idea about which gas contributes more, and even how much more.

We can also notice that other GHGs emissions are still negligible for some industries even after scaling it by a factor of 10. However for more information we must look for it separately. Which we do in Task 2. The plots from the first section gives us a clear picture about the decreasing trends in the amount of total emissions from industries and commodities. Alongside this, inferences about decreasing trend in emission values of individual Green House Gases could also be made. These plots also build up a strong foundation to detail analyses of individual gases and their trends over the years.

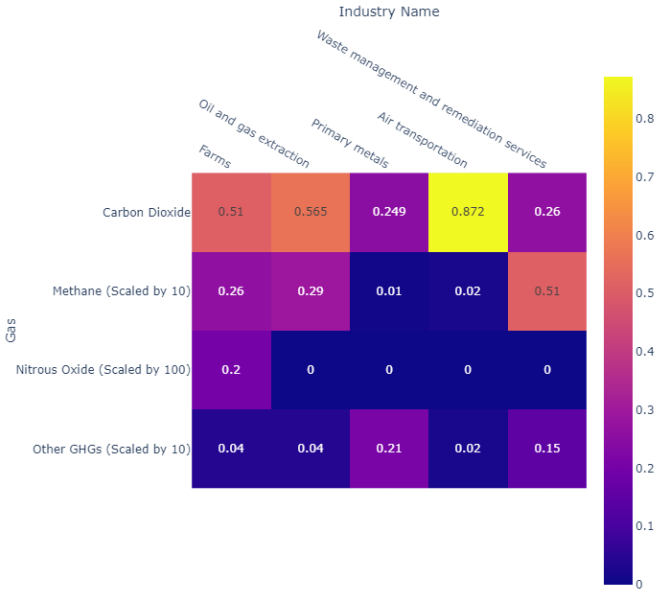


Fig. 5. Gas wise emissions of industries in 2016

B. Task-2 : Substance-wise Analysis and Contribution to Global Emissions

We have seen Emission trends for industries and commodities, now we will see how Substance-wise trends and analyze the dataset further. We begin by examining carbon dioxide, the most significant contributor to global emissions. As illustrated in Fig. 6, both the industry and commodity perspectives highlight *"Utilities"* as the dominant source of carbon dioxide emissions, largely due to the heavy reliance on fossil fuels for power generation. This sector's substantial carbon footprint is further reflected in its consistent ranking across different categorizations. Other significant contributors, such as *"Pipeline Transportation"* and *"Truck Transportation"*, also play a major role, indicating the extensive emissions associated with energy transport and logistics. The similarity in emission patterns across both views suggests that these sectors remain crucial targets for emissions reduction

initiatives, as coordinated efforts here could yield substantial benefits in reducing overall carbon dioxide levels. Moving on

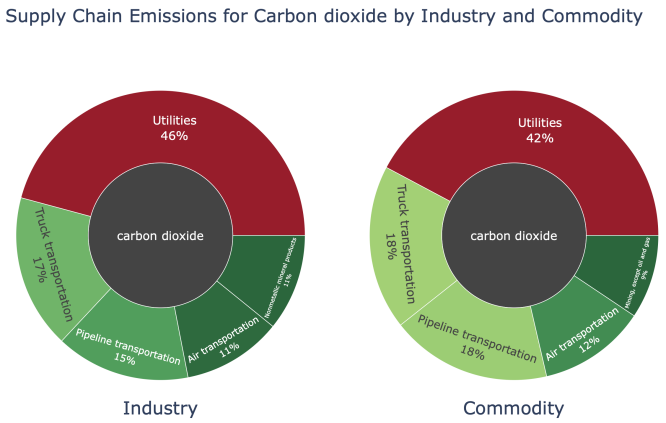


Fig. 6. Carbon dioxide split(2010-2016)

to methane, another greenhouse gas with considerable impact, Fig. 7 shows that methane emissions arise from a wider array of sources compared to carbon dioxide. *"Waste Management and Remediation Services"* emerges as a leading emitter across both the industry and commodity categories, underscoring the significant emissions linked to organic waste decomposition and landfill operations. Additionally, sectors like *"Pipeline Transportation"* and *"Farms"* are also notable contributors, reflecting the diverse origins of methane emissions, spanning waste management, energy transport, and agricultural activities. The overlapping presence of these sectors in both industry and commodity classifications implies that addressing emissions from these key areas through integrated strategies could effectively mitigate methane release into the atmosphere.

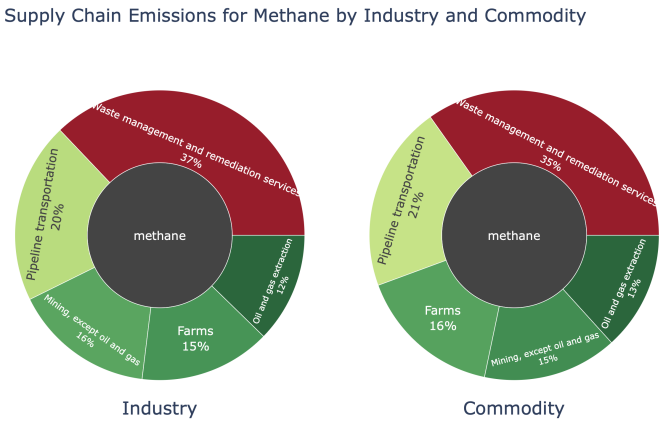


Fig. 7. Methane split(2010-2016)

Nitrous oxide, a potent greenhouse gas with a much higher global warming potential, is primarily associated with agricultural practices. Fig. 8 clearly indicates that *"Farms"* dominate nitrous oxide emissions in both the industry and commodity contexts, highlighting the critical role of activities such as fertilizer application, soil management, and livestock farming in generating nitrous oxide. The remaining contributions are

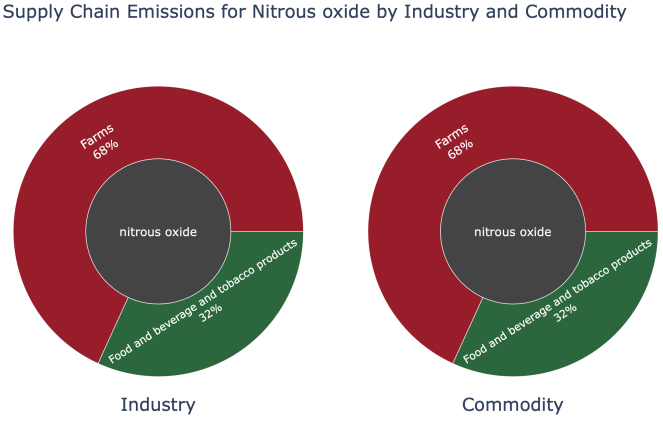


Fig. 8. Nitrous Oxide split(2010-2016)

mainly from *"Food and Beverage and Tobacco Products"*, further linking nitrous oxide emissions to agriculture-related processes. The consistent dominance of these sectors across different categorizations points to the agricultural sector as a crucial focus for reducing nitrous oxide emissions, suggesting that targeted policies in this area could achieve significant climate benefits.

The analysis of other greenhouse gases (GHGs) reveals a more diverse set of emission sources, as shown in Fig. 9. *"Water Transportation"* is a notable emitter in both industry and commodity views, likely due to emissions from marine fuel use and related activities. *"Machinery"* and *"Primary Metals"* are also key contributors, driven by specific industrial processes that release non-conventional GHGs, such as refrigerants and solvents. The similar distribution of emissions between these views suggests that tackling emissions in these sectors could have a broader impact, addressing multiple GHG types simultaneously. This highlights the importance of targeted technological innovations and regulations to reduce emissions across these varied sectors.

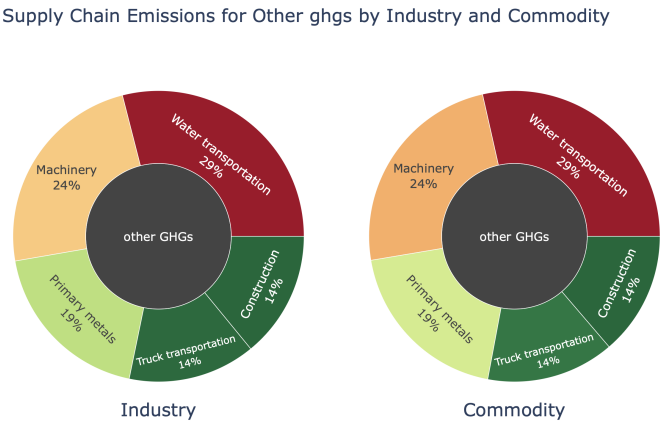


Fig. 9. Other GHG's split(2010-2016)

Overall, the substance-wise breakdown of emissions pro-

vides a detailed understanding of how different sectors contribute to each type of greenhouse gas. While carbon dioxide and methane emissions are predominantly driven by utilities, transportation, and waste management, nitrous oxide emissions are closely tied to agricultural practices. Meanwhile, other GHG emissions are linked to specific industrial activities. By understanding these distinct patterns, policymakers and stakeholders can develop more focused and effective strategies for reducing emissions, addressing the areas where they are most concentrated.

As we consider the trends across all these substances, it's crucial to understand how these emissions have evolved over time. Figure 10 provides a comprehensive view of how average supply chain emission factors, with margins, have changed from 2010 to 2016 for different substances across both commodities and industries.

Fig. 10 shows that carbon dioxide, methane, and other GHGs have similar trends for both commodities and industries, suggesting consistent yet distinct patterns. Carbon dioxide emissions show a gradual decline over the years, indicating efforts towards reducing reliance on fossil fuels and improving energy efficiency. Methane emissions, on the other hand, display a declining trend from 2010 to 2014, followed by a steady increase, which may reflect shifts in waste management practices, agricultural activities, and energy transport efficiency over time. Other GHGs experienced a notable spike in 2011 and then gradually increased, suggesting fluctuations potentially tied to specific industrial activities or regulatory changes. Meanwhile, nitrous oxide emissions remain negligible throughout the period, reinforcing its primary association with specific agricultural practices rather than broader industrial or commodity activities.

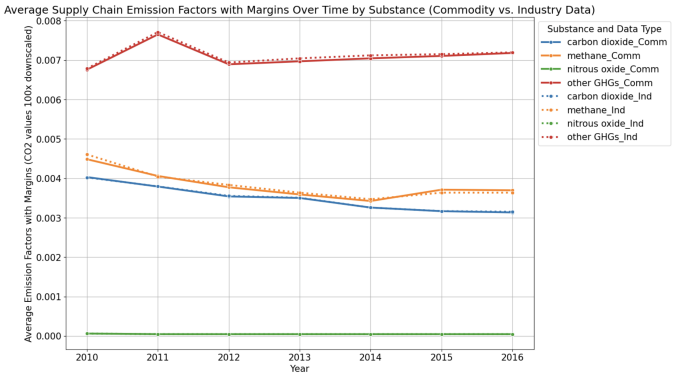


Fig. 10. Emission Trend

Overall, these trends indicate that while emissions from different substances and sectors are shifting, there is a clear need for continuous monitoring and adaptive strategies to manage and mitigate greenhouse gas emissions effectively. By understanding these evolving patterns, policymakers can better design interventions tailored to the specific needs of each sector and substance.

With these insights in mind, we now turn our focus to the quality of the data underpinning these emission estimates, examining the reliability and integrity of the information used to drive such crucial analyses.

C. Task-3: Data Quality (DQ) Scores and Their Impact on Emission Estimates

This task explores the data quality (DQ) of emission estimates for various commodities, industries and substances. The analysis focuses on understanding the reliability of the data, identifying trends, and highlighting areas where improvements can be made. Each section addresses a specific sub-question related to the overall data quality of emission estimates. Visualizations are provided to support the analysis, with each figure labelled as Fig x.

The primary columns which give us the insights of Data Quality in our dataset are - 'DQ ReliabilityScore of Factors without Margins','DQ TemporalCorrelation of Factors without Margins','DQ GeographicalCorrelation of Factors without Margins','DQ TechnologicalCorrelation of Factors without Margins','DQ DataCollection of Factors without Margins'. Other columns used in the previous sections will also be later utilized to understand their relationships. Before that, we must try to look at the distribution of data in all of the DQ columns stand alone, i.e., Univariate Analysis. Even prior to that, let's try to understand what these columns represent.

Table 3. Updated Data Quality Pedigree Matrix – Flow Indicators

		Highest score →			Lowest score →	
Indicator		**1**	**2**	**3**	**4**	**5 (default)**
Flow reliability		Verified[1] data based on measurements	Verified data based on a calculation **or** non-verified data based on measurements	Non-verified data based on a calculation	Documented estimate	Undocumented estimate
Flow Representativeness	**Temporal correlation**	Less than 3 years of difference[2]	Less than 6 years of difference	Less than 10 years of difference	Less than 15 years of difference	Age of data unknown or more than 15 years
	Geographical correlation	Data from same resolution **and** same area of study	Within one level of resolution **and** a related area of study[3]	Within two levels of resolution **and** a related area of study	Outside of two levels of resolution **but** a related area of study	From a different or unknown area of study
	Technological correlation	All technology categories[4] are equivalent	**Three** of the technology categories are equivalent	**Two** of the technology categories are equivalent	**One** of the technology categories is equivalent	**None** of the technology categories are equivalent
	Data collection methods	Representative data from >80% of the relevant market[5], over an adequate period[6]	Representative data from 60-79% of the relevant market , over an adequate period **or** representative data from >80% of the relevant market, over a shorter period of time	Representative data from 40-59% of the relevant market, over an adequate period **or** representative data from 60-79% of the relevant market, over a shorter period of time	Representative data from <40% of the relevant market, over an adequate period of time **or** representative data from 40-59% of the relevant market, over a shorter period of time	Unknown **or** data from a small number of sites **and** from shorter periods

Fig. 11. DQ Pedigree Matrix

[1] Verification may take place in several ways, e.g. by on-site checking, by recalculation, through mass balances or cross-checks with other sources. For values calculated from a mass-balance or another verification method, an independent verification method must be used in order to qualify the value as verified.
[2] Temporal difference refers to the difference between date of data generation and the date of representativeness as defined by the scope of the project
[3] A related area of study is defined by the user and should be documented in the geographical metadata. The relationship established in the metadata of the unit process should be consistently applied to all flows within the unit process. Default relationship is established as within the same hierarchy of political boundaries (e.g. Denver is within Colorado, is within the USA, is within North America)
[4] Technology categories are process design, operating conditions, material quality, and process scale.
[5] The relevant market should be documented in the DQG. The default relevant market is measured in production units. If the relevant market is determined using other units, this should be documented in the DQG. The relevant market established in the metadata should be consistently applied to all flows within the unit process.
[6] Adequate time period can be evaluated as a time period long enough to even out normal fluctuations. The default time period is 1 year, except for emerging technologies (2-6 months) or agricultural projects >3 years.

Take a look at this table. This table presents a Data Quality Pedigree Matrix used to evaluate the quality of data based on

several key indicators. These indicators help assess the reliability and representativeness of flow data for various applications, particularly in lifecycle analysis or similar studies. This Matrix would be very helpful for us to give conclusions about the visualizations created henceforth. Source to this is Here

I would first like to put an emphasis on the distribution of the columns – 'DQ GeographicalCorrelation of Factors without Margins' and 'DQ DataCollection of Factors without Margins'. These column seems to have all values equal to 1, which means all the commodities have the Data that comes from the same resolution and area as the study and representative data is from over 80% of the relevant market, gathered over an adequate time period. Which means that all this data is recorded in the same area/city/state/country and is relevant. So there cannot be any biases related to the geographical locations of the industries. A pie chart for this would have the whole circle covered with the same color as there is no other value apart from '1'. Let us take a look at the other distributions. The following plot includes Industry and commodity – pairwise plots for each of the different DQ related columns. The pie charts for DQ Reliability Score

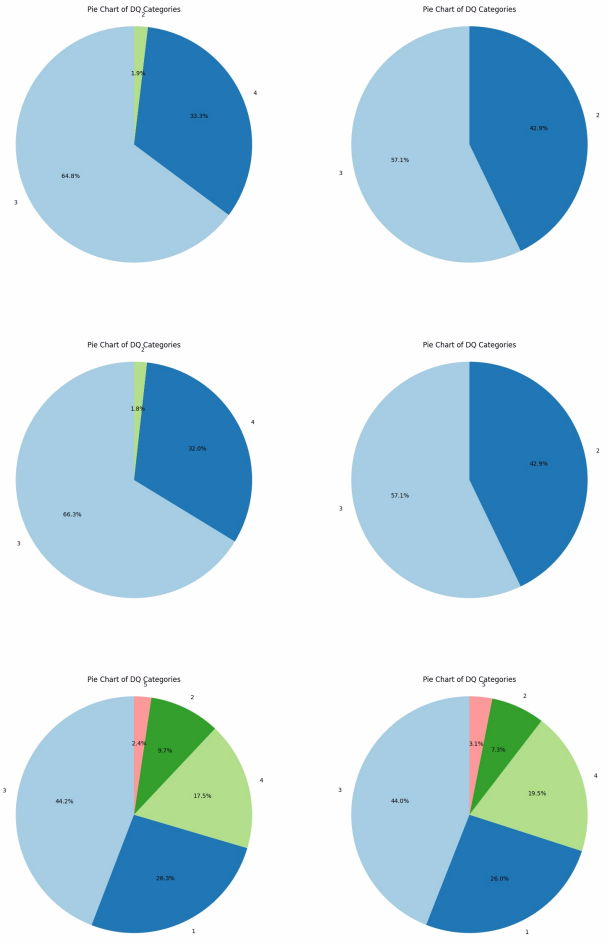


Fig. 12. Pairwise Pie Chart of Commodity-Industry Data

reveal that the majority of data falls under categories 3 and 4, indicating heavy reliance on documented estimates or non-verified calculations. Over 30% of the data is based on less precise methods, suggesting a need for better verification processes. A small portion of category 2 indicates some verified data, but the absence of category 1 shows a lack of highly reliable, measured data across both commodities and industries. The DQ TemporalCorrelation charts show that most data is moderately recent, with a majority falling into categories 2 and 3 (up to 10 years old). This suggests the data is somewhat outdated but still usable. With over 50% of the data falling under category 3, concerns about its recency arise, and the absence of category 1 highlights the need for more frequent updates to maintain relevance. The two pie charts represent the Technological Correlation of data quality for Commodity and Industry categories. Both charts reveal that the most significant portion of the data is rated as a "3," indicating that the technological correlation in these data sets is moderately aligned with the study's technological framework. This suggests that, while the data reflects some aspects of the required technological conditions, there are notable discrepancies or gaps that could impact the accuracy and reliability of any conclusions drawn from these data sets.

In addition, both charts show a significant proportion of the data rated as "1" and "2," signifying a strong alignment in some areas, but also a presence of "4" and "5" ratings, indicating poor alignment for certain technological categories. The slightly higher percentages of lower ratings (4 and 5) in the Commodity plot as compared to the Industry plot suggest that the technological data quality for commodities may be slightly less reliable overall. This distribution indicates that while both sectors have a substantial amount of data closely aligned with technological requirements, there are critical areas where the data does not fully meet the necessary technological standards, potentially leading to less robust analytical outcomes in those specific areas. Now that we are done with the univariate analysis, let us move forward and try to answer the sub questions mentioned earlier. Here's heatmaps showing the average DQ score of different commodities for different Emission substances. Please note the fact that only the data in the year 2016 (most recent- to include most of the new indutries) was used to plot this for clarity of visualization and for it to be meaningful.

The heatmaps presented for different greenhouse gases—carbon dioxide, methane, nitrous oxide, and other GHGs—illustrate the Data Quality (DQ) Scores for various commodities. Lower DQ scores indicate better data quality, representing more reliable and representative data for assessing emissions, while higher scores suggest greater uncertainty or gaps in data quality. Each heatmap provides a visual summary of the strengths and weaknesses in data quality for different commodities and substances.

For carbon dioxide, the heatmap reveals a range of DQ scores, with certain commodities such as *"Accommodation"* and *"Utilities"* demonstrating lower scores. These lower scores reflect high data reliability and consistency, possibly

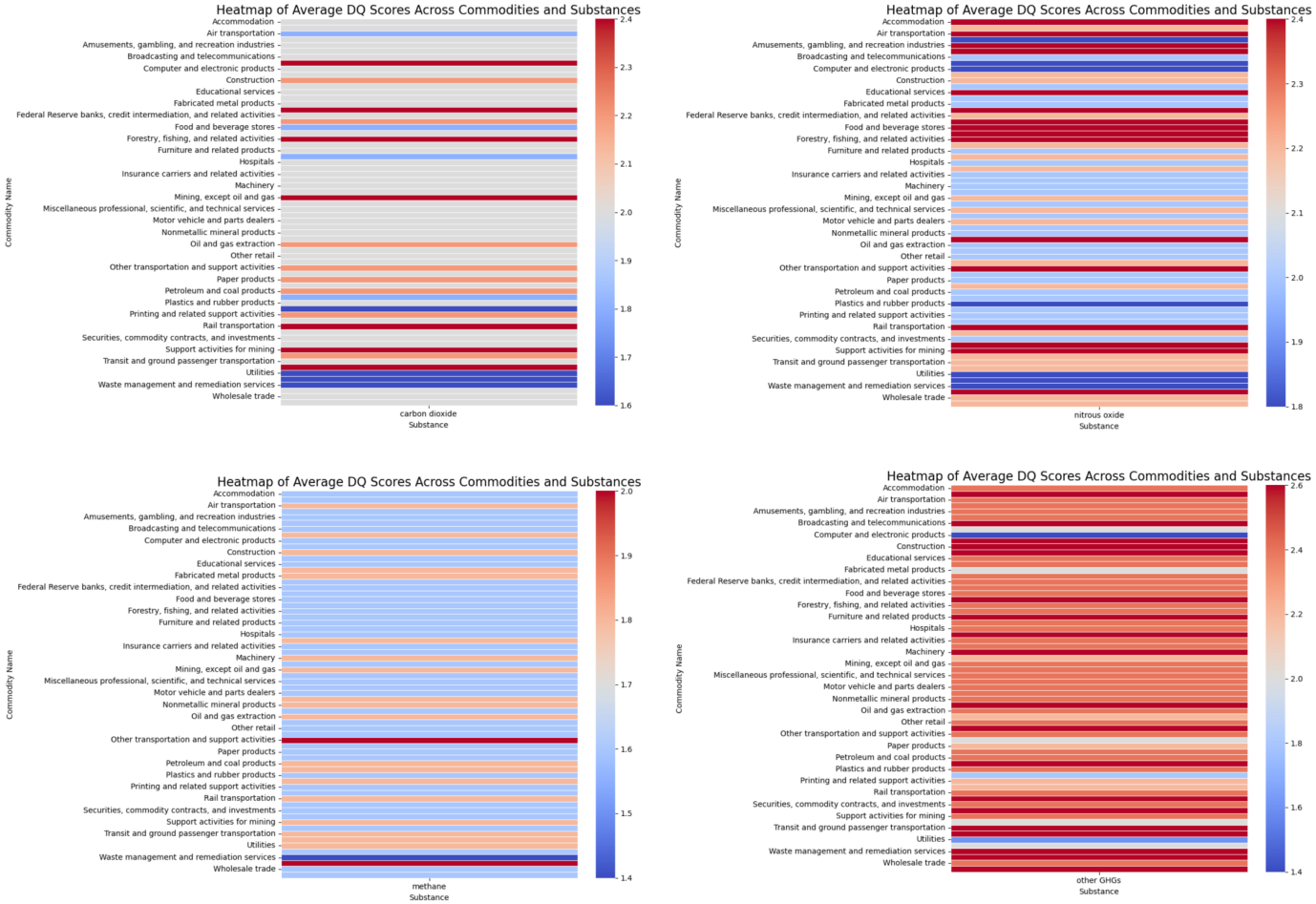


Fig. 13. Heatmap of Average DQ Score versus Emmision substance

due to more standardized methods of measuring carbon dioxide emissions in these sectors. In contrast, sectors like *"Air transportation"* and *"Waste management and remediation services"* show higher DQ scores, suggesting the presence of more significant uncertainties or less reliable data collection methods, making the data for these sectors less robust.

The heatmap for methane shows a trend where most commodities have higher DQ scores, reflecting challenges in obtaining accurate and reliable methane emissions data. However, some commodities like *"Rail transportation"* and *"Utilities"* display relatively lower scores, indicating more dependable data sources or better-established data collection practices for these commodities. On the other hand, sectors like *"Federal Reserve banks, credit intermediation, and related activities"* and *"Broadcasting and telecommunications"* exhibit higher DQ scores, highlighting potential data quality gaps that could affect the accuracy of methane emission assessments.

For nitrous oxide, the heatmap presents a mixed pattern. Several commodities, such as *"Insurance carriers and related activities"* and *"Other retail"*, have lower DQ scores, suggesting a higher quality of data for nitrous oxide emissions in these sectors. This indicates more reliable data collection and representativeness. Conversely, commodities like *"Food and beverage stores"* and *"Furniture and related products"* show higher DQ scores, suggesting that the data quality for nitrous oxide emissions in these sectors might need further improvement to reduce uncertainties.

The heatmap for other GHGs reveals significant variability in DQ scores across commodities. For example, sectors such as *"Printing and related support activities"* and *"Nonmetallic mineral products"* show lower scores, indicating strong data reliability and representativeness. However, higher DQ scores for commodities like *"Amusements, gambling, and recreation industries"* and *"Air transportation"* point to potential inconsistencies or uncertainties in data reporting, which could lead to less accurate assessments for these gases.

Overall, these heatmaps provide a comprehensive view of the data quality landscape across different commodities and substances, emphasizing where data is most reliable and where there are gaps. The varying DQ scores across these sectors and substances highlight the need for more targeted efforts to improve data collection practices and ensure more reliable data for emission assessments. Addressing these gaps is crucial for formulating effective policies and interventions for managing greenhouse gas emissions more accurately.

Here's another visualization - a divergent bar chart to show

the specific commodities with the best and worst quality of data.

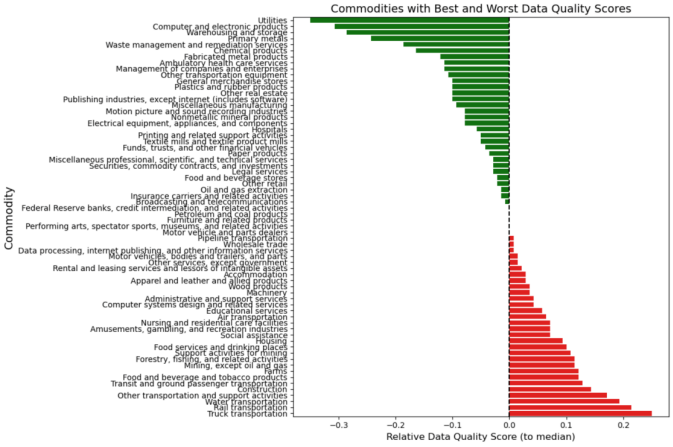
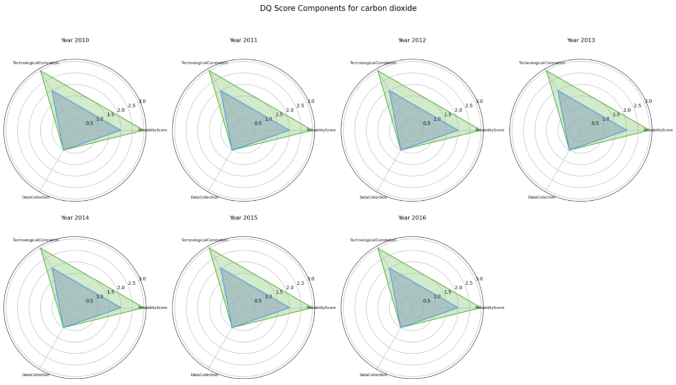


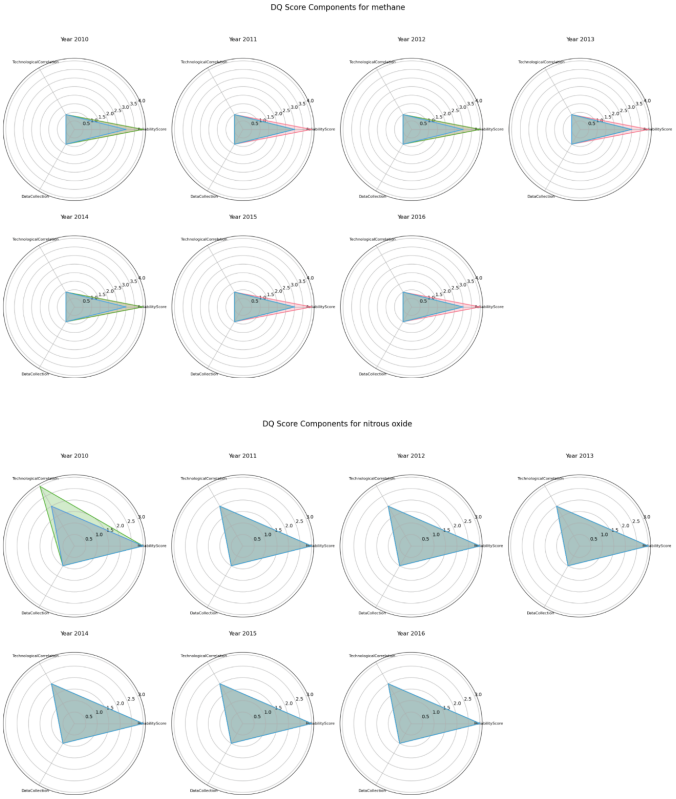
Fig. 14. Divergent Bar Graph to find out specific commodities

The divergent bar chart displays the relative Data Quality (DQ) scores for various commodities, distinguishing those with the best (lower scores) and worst (higher scores) data quality. Commodities like *"Utilities," "Computer and electronic products,"* and *"Warehousing and storage"* have lower DQ scores, indicating highly reliable and representative data for emission estimates. These sectors likely benefit from robust data collection and verification methods. In contrast, commodities such as *"Truck transportation," "Rail transportation,"* and *"Water transportation"* have higher DQ scores, suggesting significant data quality issues due to factors like diverse sources, complex supply chains, or inadequate data coverage. Improving data quality in these areas would enhance the reliability of emission assessments.

Now that we know that the commodities with the best quality data are - *"Utilities," "Computer and electronic products,"* and *"Warehousing and storage"*, we can try to analyse a temporal trend in their individual DQ contributors. As I mentioned earlier, Geographical and DataCollection DQ scores will not be counted as they are uniform. The radar plots for



different greenhouse gases—carbon dioxide, methane, nitrous oxide, and other GHGs—illustrate the changes in Data Quality (DQ) score components over time from 2010 to 2016. Each



plot breaks down three key components: *Technological Correlation*, *Reliability Score*, and *Data Collection*, providing a comprehensive view of the data quality landscape for each substance.

For carbon dioxide, the plots show a relatively balanced distribution across all three components, with *Technological Correlation* generally achieving lower (better) scores compared to *Reliability Score* and *Data Collection*. This suggests that data quality related to technological aspects, such as methods and technologies used for tracking carbon dioxide, is more robust. Meanwhile, *Reliability Score* and *Data Collection* have slightly higher scores, indicating areas where further improvement is possible to enhance the overall data accuracy.

In contrast, methane displays a noticeably higher range across all components, particularly for the *Reliability Score*. This suggests significant challenges in capturing reliable methane emissions data, likely due to the more diffuse and varied sources of methane compared to carbon dioxide. The consistently high scores from 2010 to 2016 imply that data quality issues related to methane have remained relatively unchanged over the years, indicating a need for enhanced data collection and reporting methods.

For nitrous oxide, the plots show relatively low scores across all three components, with minimal fluctuations. This uniformity suggests that the data quality for nitrous oxide has been relatively stable and reliable, particularly in terms of *Technological Correlation* and *Data Collection*. However, there is still room for improvement in further lowering the *Reliability Score* to achieve even more accurate data.

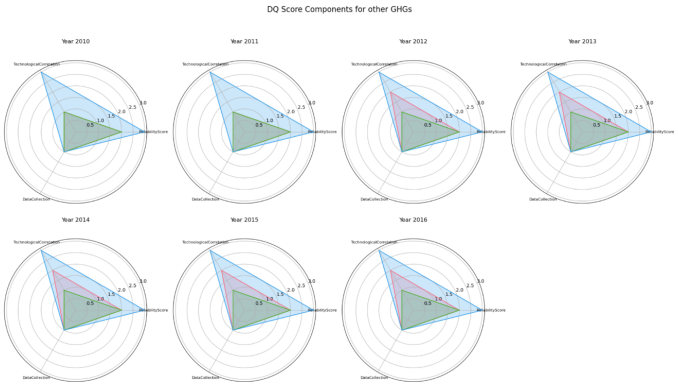


Fig. 15. Radar Plots for chronological analysis of DQ contributors

The plots for other GHGs reveal more variability across the years. *Technological Correlation* generally achieves lower scores, indicating strong data quality in terms of technological aspects. However, the *Reliability Score* and *Data Collection* components exhibit more variation, with certain years reflecting higher scores. This variability suggests that while technology for monitoring these gases may be in place, the reliability and consistency of the data could be compromised due to factors like less frequent reporting or complex data integration challenges.

Overall, these radar plots highlight the need for focused efforts to further lower DQ scores—particularly for methane and other GHGs—to enhance the reliability and accuracy of emission estimates. Improved data collection methods, more consistent reporting, and better technological integration are essential for achieving high-quality emission data across all greenhouse gases.

We have come this far and gained some useful insights on the basis of DQ columns standalone. Let us now look at something like a multivariate analysis of these DQ contributors with the real important columns, i.e., emission factors. Here's two Hexagonal 2-D heatmap which will help us to gain insights into their relationships. The blue one is for 'Margins of Supply Chain' and the red one is for 'Supply Chain emission factors without Margins'.

This blue hexbin plot illustrates the relationship between the Average DQ Score and Margins of Supply Chain Emission Factors. The plot reveals that lower DQ scores (indicating better data quality) generally align with lower margins in emission factors, suggesting that higher-quality data tends to correspond with more precise and consistent emission estimates.

Clusters of darker hexagons around DQ scores of approximately 2.0 to 2.2 and lower margins indicate a concentration of commodities where the data quality is reasonably good, and the variability in emission factors is controlled. Conversely, as the DQ scores increase beyond 2.4, we observe a broader spread and higher margins in emission factors, highlighting the impact of less reliable data on the uncertainty of emission estimates.

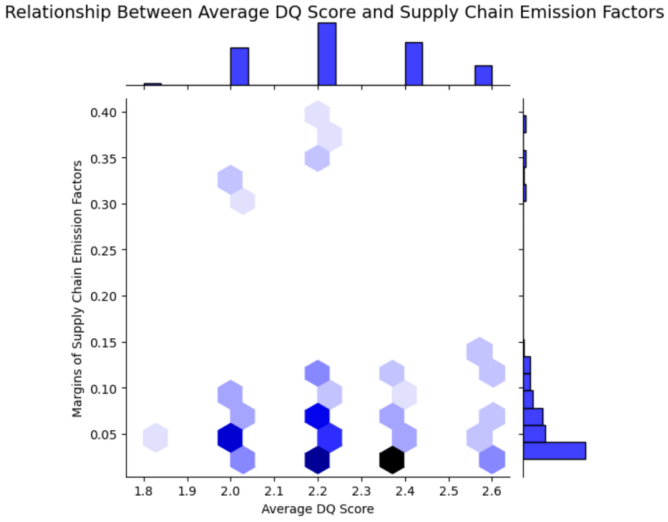


Fig. 16. HeatMap to understand relation b/w DQ contributors and emission factors

The marginal histograms further emphasize this pattern, showing a higher density of commodities with better DQ scores and lower emission factor margins. This reinforces the importance of improving data quality to achieve more accurate and reliable emissions reporting in supply chain analysis.

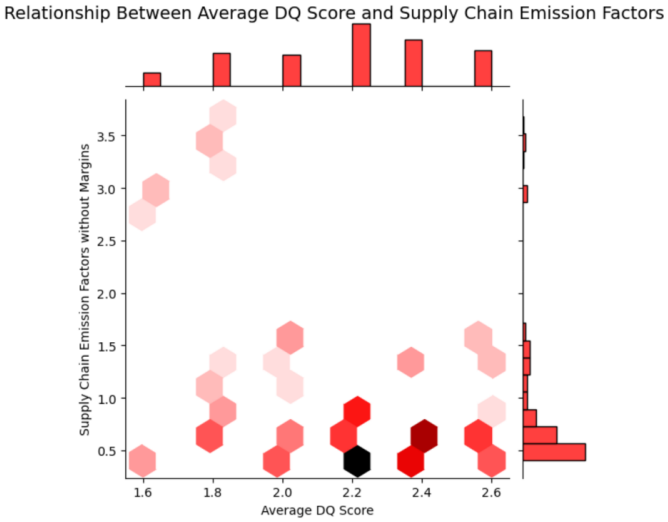


Fig. 17. HeatMap to understand relation b/w DQ contributors and emission factors

This red hexbin plot illustrates the relationship between the Average DQ Score and Supply Chain Emission Factors without Margins. The plot reveals a general pattern where commodities with lower DQ scores (better data quality) tend to have lower supply chain emission factors. This suggests that higher data quality is associated with more accurate and potentially lower emission estimates.

Clusters of darker hexagons appear around DQ scores between 1.8 and 2.2, with lower emission factors, indicating that commodities with better data quality tend to have less

variability in emission estimates. As DQ scores increase beyond 2.4, there is a noticeable spread in emission factors, implying greater uncertainty or inconsistency in data quality that could lead to higher reported emissions.

The marginal histograms further emphasize this pattern, showing a higher density of commodities with better DQ scores on the left side, corresponding to lower emission factors. This relationship underscores the importance of improving data quality to achieve more precise and reliable emissions data for supply chain analysis.

We are almost at the end of this task and the last thing which remains is to figure out whether the DQ contributors affect one another. Can Technological DQ score improve ReliabilityScore DQ? Or can the Temporal one bring changes? Such questions will be answered now.

Here are three sunburst plots to try and find out some relationship between categorical columns. The only columns which will be useful are 'DQ ReliabilityScore of Factors without Margins','DQ TemporalCorrelation of Factors without Margins','DQ TechnologicalCorrelation of Factors without Margins'. That is the reason why we have 3 sunbursts for this Bivariate analysis.

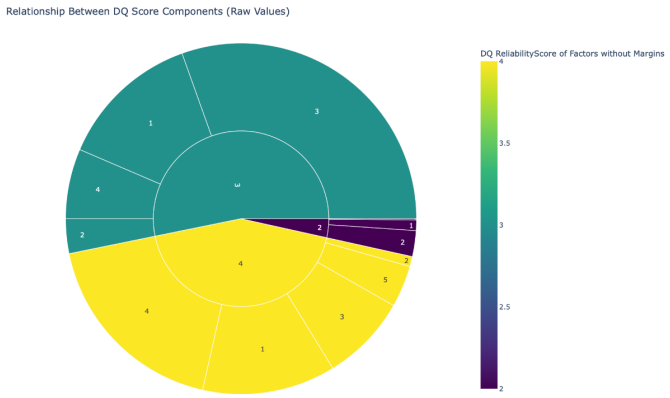


Fig. 18. Sunburst - Reliability vs Technological

The sunburst chart visualizes the relationship between different components of Data Quality (DQ) scores for emission factors without margins. The inner circle represents the DQ Reliability Score, while the outer layers expand into other DQ components, such as Technological Correlation.

Lower DQ scores (closer to 1) indicate higher data quality and are concentrated in smaller sections of the inner circle, shaded in darker colors. These segments represent data that is verified and closely aligned with the study's technological conditions. In contrast, higher DQ scores (closer to 5) are represented by larger sections in lighter colors, reflecting less reliable data based on estimates or mismatched technological conditions.

The chart highlights the interconnectedness of different data quality aspects, showing how reliability in one area (e.g., technological correlation) can impact overall data quality. This visualization is helpful for identifying areas where improvements are needed to ensure more robust and accurate data for emission assessments.

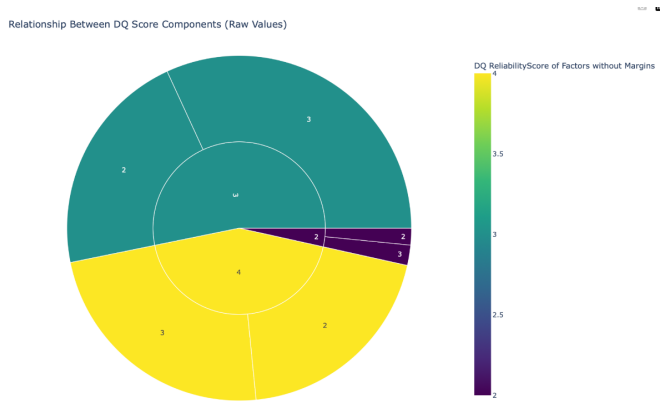


Fig. 19. Sunburst - Reliability vs Temporal

The updated sunburst chart displays the relationship between different Data Quality (DQ) Score Components for emission factors without margins. The inner circle represents the DQ Reliability Score, while the outer layers depict additional DQ components such as Temporal Correlation.

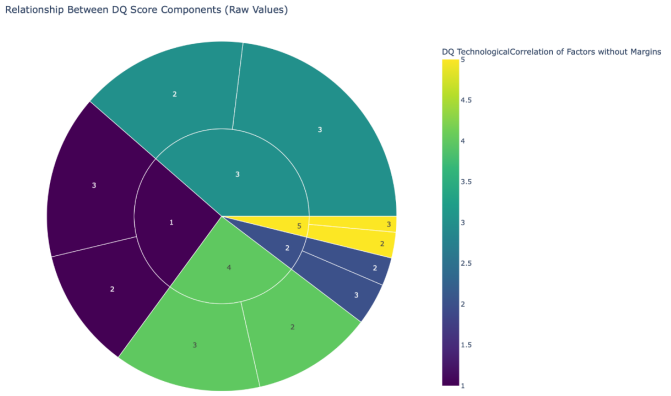


Fig. 20. Sunburst - Technological vs Temporal

The sunburst chart provides an in-depth visualization of the relationship between different Data Quality (DQ) Score Components for emission factors without margins. The inner circle represents the DQ Technological Correlation, while the outer layers depict the Temporal Correlation of the data.By examining the distribution across these layers, we can discern how both technological alignment and the timeliness of data affect overall data robustness. This chart helps identify specific areas that may need targeted improvements to enhance the quality and applicability of data for accurate emission assessments.

The three sunburst charts collectively illustrate the interplay between different Data Quality (DQ) Score Components—such as Reliability, Technological Correlation, and Temporal Correlation—and how they influence the overall quality of emission data. Lower DQ scores, represented in

darker shades, consistently show higher data reliability and alignment with study conditions. This indicates that when all components are well-correlated, the data becomes more robust and suitable for accurate emissions analysis.

However, as the scores increase toward 5, the lighter sections of the charts reveal areas of concern. Higher DQ scores highlight potential gaps or inconsistencies in data quality, such as outdated information or poor technological relevance. These gaps may affect the accuracy of lifecycle assessments or policy decisions based on the emission data.

Overall, the sunburst charts underscore the need for a balanced improvement across all DQ components. By focusing on enhancing technological alignment, timeliness, and reliability, stakeholders can ensure that the data used for emission estimates is both reliable and actionable, paving the way for more informed environmental strategies and policies.

IV. Member-wise Contributions

- Task-1 : Krish Patel
- Task-2 : Vansh Sinha
- Task-3 : Ashirwad Mishra