Group9_Project1

Anika Kathuria, Yixuan Chen, Nan Xiao, Obiora Okeke

2025-02-18

I Introduction and dataset description

To study the relationship between supply chain greenhouse gas (GHG) emission factors and various industries, as well as to analyze and measure carbon emissions within the supply chain through data analysis, we have identified this dataset. It adopts the North American Industry Classification System (NAICS) to classify industries, covering sectors such as agriculture, manufacturing, energy, transportation, and services. Each industry has a corresponding GHG emission factor, with three categories of data: without Margins, Margins, and with Margins. These categories allow for further research on the relationship between supply chain carbon emissions and economic benefits.

This dataset contains 1,016 rows and 8 columns, covering information on NAICS Code (North American Industry Classification System), Greenhouse Gas (GHG) emissions, Supply Chain Emission Factors, Unit, and USEEIO (U.S. Environmental Economic Input-Output Model). Each row represents the greenhouse gas emissions data of an industry (NAICS Code) within the supply chain. The 8 columns are as follows:

- NAICS Code (Column 1): The North American Industry Classification System (NAICS) code used to identify different industries.
- NAICS Title (Column 2): The name of the industry.
- GHG (Column 3): The type of greenhouse gas, which may represent the total emissions of all greenhouse gases or specific types.
- Unit (Column 4): The unit of emission factors, such as "kg CO_2e/2022 USD, purchaser price," representing the greenhouse gas emissions generated per unit of economic activity.
- Supply Chain Emission Factors (Columns 5, 6, 7): These columns include different types of supply chain emission factors, possibly representing supply chain emissions excluding profit margins, profit margins, and total supply chain emissions.
- USEEIO Code (Column 8): The industry code in the USEEIO model.

Columns 5, 6, and 7, which represent the three Supply Chain Emission Factors, are numerical variables with different meanings.

- Supply Chain Emission Factors without Margins: This column represents the carbon emission factor of upstream activities in the supply chain, excluding profit or markup factors. It measures the greenhouse gas emissions (typically in kg CO_2e) generated by an industry to meet one dollar of final demand (calculated in 2022 USD) during the production process. If the Supply Chain Emission Factors without Margins is 0.488, it means that for every one dollar of output in this industry, 0.488 kg CO_2e of supply chain carbon emissions are produced.
- Margins of Supply Chain Emission Factors: This column represents the additional carbon emission factor caused by profit, markups (margins), and distribution costs (such as transportation and sales) in the supply chain. For example, price increases at different stages (retail, wholesale, etc.) may indirectly lead to additional carbon emissions. If the Margins of Supply Chain Emission Factors is 0.044, it means that the additional carbon emissions caused by profit and associated costs in this industry amount to 0.044 kg CO_2e.
- Supply Chain Emission Factors with Margins: This column represents the total supply chain emission factor, which includes emissions from production, profit, and additional costs (such as distribution and sales), reflecting the total greenhouse gas emissions associated with the supply chain.

From the explanations above, it is evident that these three indicators provide a clear representation of carbon emissions across various industries. They collectively illustrate the environmental impact of different industries from multiple perspectives, including the production phase, profit, distribution, and the complete supply chain. Analyzing these factors can aid in optimizing supply chains by identifying industries with the highest environmental impact during production. Additionally, they offer insights into the overall impact of final consumption on the environment, supporting the evaluation of an industry's full life cycle carbon footprint and informing the development of carbon neutrality policies.

II Data acquisition methodology

The "Supply Chain Greenhouse Gas Emission Factors v1.3 by NAICS-6" dataset was developed by the U.S. Environmental Protection Agency (EPA), specifically by its Office of Research and Development (ORD) in 2024. This dataset provides greenhouse gas (GHG) emission factors for 1,016 U.S. commodities, categorized according to the 2017 North American Industry Classification System (NAICS) at the 6-digit level. This dataset was created at the request of the General Services Administration (GSA) to aid in annual GHG reporting for federal agencies. The dataset is publicly accessible and was obtained from Data.gov, a federal open data platform. The link to the dataset is: https://catalog.data.gov/dataset/supply-chain-greenhouse-gas-emission-factors-v1-3-by-naics-6

The originally obtained data is relatively clean, leaving little room for data processing and cleaning. Therefore, Python was used to shuffle the raw data and artificially introduce noise, resulting in

Table 1: First 6 Rows of the Dataset

X2017.NAICS.Code	X2017.NAICS.Title	GHG	Unit	Supply.Chain.Emission.Factors.without.Margins	Margins.of.Supply.Chain.Emission.Factors	Supply.Chain.Emission.Factors.with.Margins	Reference.USEEIO.Code	X2017.NAICS.Title.1
425120	Wholesale Trade Agents and Brokers	All GHGs	mg CO2e/kg	0.092	0.000	0.092	425000	NA
336419	Other Guided Missile and Space Vehicle Parts and Auxiliary Equipment Manufacturing	All GHGs#	kg CO2e/2022 USD, purchaser pric&	0.286	0.007	0.294	33641A	NA
337124	Metal Household Furniture Manufacturing	All GHGs#	ton CO2e	0.168	0.060	0.230	33712N	NA
337127	Institutional Furniture Manufacturin%	All GHGs%	ton CO2e	0.208	0.038	0.246	33712#	NA
561440	Collection Agencies%	All GHGs#	ton CO2e#	0.111	0.000	0.111	561400	NA
333517	Machine Tool Manufacturing	All GHGs	kg CO2e/2022 USD, purchaser price	0.176	0.000	0.199	33351#	NA

issues such as garbled characters, inconsistent units, changes in numerical precision, the presence of outliers and extreme values, duplicate data, misaligned data, and missing values. The file name for the code used to make the data more messy is: mess.R. The messed dataset is as follows.

III Cleaning and preprocessing steps

The issues in the dataset have been addressed through a cleaning process. By initially observing the data structure, several significant problems were identified, such as garbled characters, missing values, and structural inconsistencies. The cleaning process was carried out systematically, addressing column-related issues first, followed by row-related issues.

1. The dataset originally contained 9 columns, but the ninth column consisted entirely of NA values and had a duplicate column name with Column 2. Since this column was deemed redundant and invalid, it was removed.

2. The names in Column 1 and Column 2 started with an “X,” which had no meaningful significance. This special character was removed.

3. Column 1 contained numerical codes that should be six-digit integers. To standardize this, all values in Column 1 were converted to their absolute values, rows with values exceeding 1,000,000 were removed, and the remaining values were stored as integers.

4. Columns 2 and 3 contained string data without special characters. Any unnecessary special characters were removed. Additionally, since the dataset recorded greenhouse gas types as “All GHGs,” all values in Column 3 were standardized to “All GHG.”

5. Column 4 represented units. The dataset exhibited inconsistencies in units, but after examining the data, it was found that all three Supply Chain Emission Factors shared a similar scale. Therefore, after further verification, all values were unified to “kg CO2e/2022 USD, purchaser price” to ensure consistency.

6. Columns 5, 6, and 7 contained numerical data, but inconsistencies in formatting and data errors were observed. To address this, all values were converted to their absolute values (since carbon emissions must be positive), outliers were removed, and decimal points were standardized for consistency and clarity.

Table 2: First 6 Rows of the Dataset

X2017.NAICS.Code	X2017.NAICS.Title	GHG	Unit	Supply.Chain.Emission.Factors.without.Margins	Margins.of.Supply.Chain.Emission.Factors	Supply.Chain.Emission.Factors.with.Margins	Reference.USEEIO.Code
111,110	Soybean Farming	All GHGs	kg CO2e/2022 USD, purchaser price	0.488	0.000	0.532	11110A
111,120	Oilseed (except Soybean) Farming	All GHGs	kg CO2e/2022 USD, purchaser price	0.488	0.044	0.532	11110A
111,150	Corn Farming	All GHGs	kg CO2e/2022 USD, purchaser price	0.809	0.040	0.850	11110B
111,160	Rice Farming	All GHGs	kg CO2e/2022 USD, purchaser price	0.809	0.000	0.848	11110B
111,191	Oilseed and Grain Combination Farming	All GHGs	kg CO2e/2022 USD, purchaser price	0.809	0.000	0.850	11110B
111,199	All Other Grain Farming	All GHGs	kg CO2e/2022 USD, purchaser price	0.809	0.040	0.848	11110B

7. Column 8 contained categorical codes. Upon investigation, the USEEIO code was found to be a five-character alphanumeric code. Any garbled characters were removed, and missing digits were filled with leading zeros to maintain uniformity.

8. Duplicate rows were identified and removed to resolve data redundancy issues.

9. The entire dataset was sorted in ascending order based on the NAICS code, making the cleaned dataset more visually organized and readable.

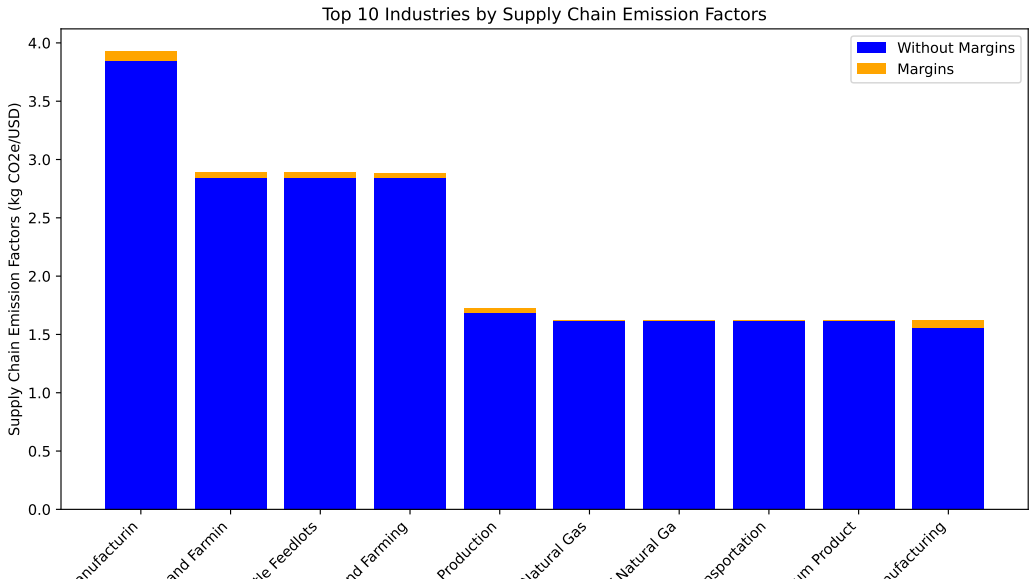
The cleaned and structured dataset is presented below.

IV Exploratory Data Analysis (EDA)

Before conducting EDA, we first need to clarify the research objective. In this dataset, we aim to study the relationship between supply chain greenhouse gas (GHG) emission factors and various industries, as well as whether there are internal correlations between different types of emission factors and whether an economic perspective can provide any insights.

1. Top 10 industries with the highest Supply Chain Emission Factors

```
## <matplotlib.legend.Legend object at 0x000001E1E9277B60>
```



This chart presents the top 10 industries with the highest supply chain emission factors, using a stacked bar chart to compare emission factors with and without margins. The height of each bar represents the amount of CO_2 emissions produced during the manufacturing process of each industry.

From the results, it is evident that among the top ten industries, the Cement Manufacturing industry has the highest supply chain emission factor, reaching nearly 4.0 kg CO_2e/USD, significantly higher than other industries. The cattle ranching and dairy-related industries (such as beef cattle ranching, cattle feedlots, and dairy cattle production) also exhibit relatively high emission factors, all approaching 3.0 kg CO_2e/USD. The pipeline transportation of petroleum and natural gas industries have relatively lower emission factors, ranging between 1.5 - 2.0 kg CO_2e/USD. The marginal emission factors (represented by the orange segments) are relatively small, indicating that for these industries, the majority of supply chain emissions are direct, with marginal contributions being less significant.

Therefore, the conclusion can be drawn that supply chain carbon emissions are primarily driven by energy-intensive industries, particularly cement manufacturing and livestock-related industries. While petroleum and natural gas transportation also contribute to supply chain emissions, their emission intensity per unit of economic output is lower compared to cement and livestock industries. Supply chain optimization and emission reduction policies should prioritize high-emission industries, such as cement manufacturing and livestock farming, to effectively reduce the overall carbon footprint.

##2. Analysis of the distribution of supply chain emission factors across industries

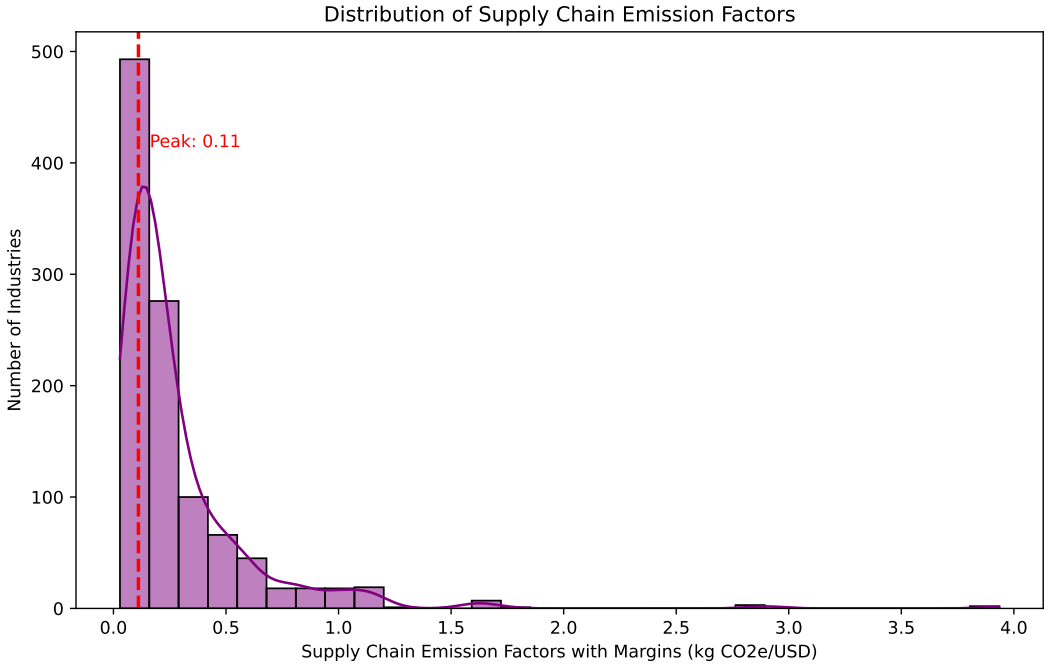
```
## <Figure size 1000x600 with 0 Axes>

## <Axes: xlabel='Supply.Chain.Emission.Factors.with.Margins', ylabel='Count'>

## <matplotlib.lines.Line2D object at 0x000001E1E9894560>
```

```
## Text(0.161, 414.12, 'Peak: 0.11')
```

```
## Text(0.5, 0, 'Supply Chain Emission Factors with Margins (kg CO2e/USD)')
```

```
## Text(0, 0.5, 'Number of Industries')
```

```
## Text(0.5, 1.0, 'Distribution of Supply Chain Emission Factors')
```



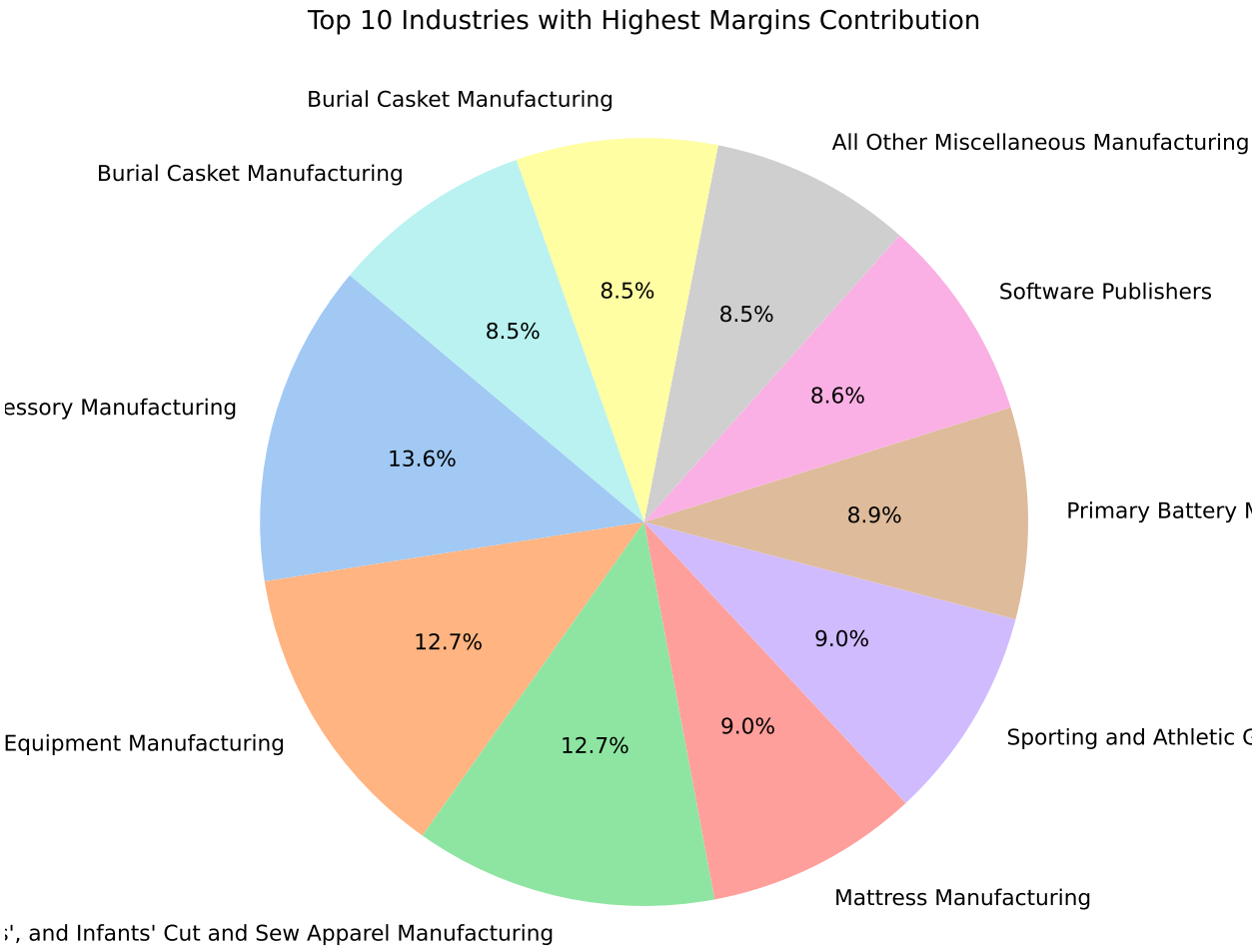
Plot a histogram to illustrate the frequency distribution of emission factors across industries. The results show that the data distribution is right-skewed (positively skewed), meaning that most industries have relatively low supply chain emission factors, while a small number of industries exhibit significantly higher emission factors.

This indicates that the carbon emission intensity is relatively low for most industries, but certain industries, such as cement manufacturing and livestock farming, have extremely high supply chain emissions. The most frequent emission factor value is 0.11 kg CO_2e/USD, suggesting that the majority of industries have supply chain emission factors concentrated around 0.1. This also implies that lower carbon emission factors are more common within supply chains.

Additionally, the histogram exhibits a long-tail effect, with the maximum value on the X-axis approaching 4.0 kg CO_2e/USD. Although only a few industries fall into this range, their emission intensity is significantly higher than the average. These high-emission industries likely make a substantial contribution to overall carbon emissions and should be key targets for future emission reduction policies.

##3. Top 10 industries of margins of Supply Chain Emission Factors

```
## <Figure size 800x800 with 0 Axes>

## ([<matplotlib.patches.Wedge object at 0x000001E1E96E4980>, <matplotlib.patches.Wedge

## Text(0.5, 1.0, 'Top 10 Industries with Highest Margins Contribution')
```



In addition to presenting the top 10 industries with the highest Supply Chain Emission Factors, we also analyzed the top 10 industries with the highest Margins of Supply Chain Emission Factors. Each sector in the pie chart represents an industry, with its size indicating the industry's share of marginal contribution.

The difference between Margins of Supply Chain Emission Factors and Supply Chain Emission Factors is that margins focus on the impact of the value-added portion on carbon emissions, primarily considering factors such as retail, wholesale, and transportation. The influencing factors include price premiums, brand value, and logistics costs.

From the pie chart results, it is evident that the industry with the highest Margins of Supply Chain Emission Factors is Cutting Tool and Machine Tool Accessory Manufacturing, accounting for 13.6%. This is followed by Women's, Girls', and Infants' Cut and Sew Apparel Manufacturing and Lawn and Garden Tractor and Home Lawn and Garden Equipment Manufacturing, both contributing 12.7%. These industries typically have high product value-added components and may involve complex supply chains and production processes.

The top 10 industries mainly revolve around apparel, tools, and garden machinery, with the presence of the software industry as well. This suggests potential opportunities for advancements in green computing and data center optimization in the future.

##4. Comparison of Emissions with and Without Margins

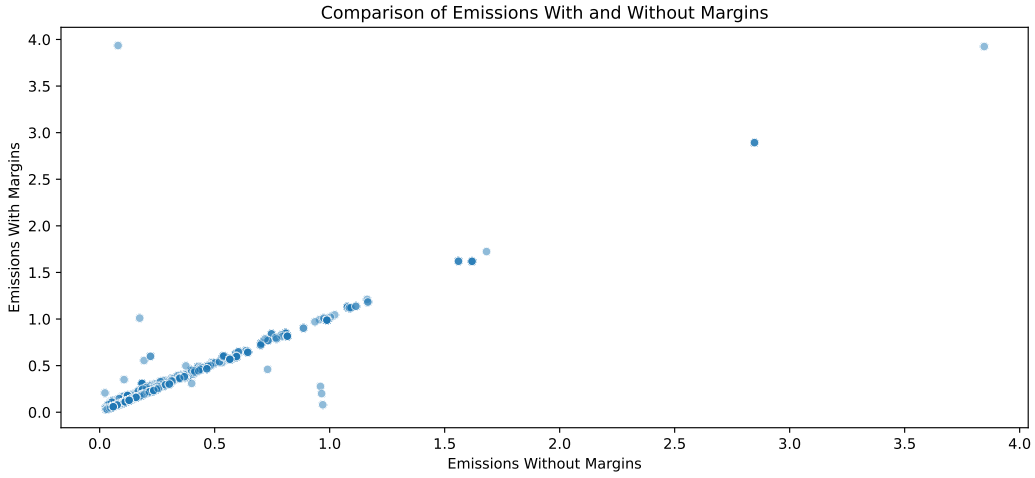
```
## <Figure size 1200x500 with 0 Axes>
```

```
## <Axes: xlabel='Supply.Chain.Emission.Factors.without.Margins', ylabel='Supply.Chain.E
```

```
## Text(0.5, 0, 'Emissions Without Margins')
```

```
## Text(0, 0.5, 'Emissions With Margins')
```

```
## Text(0.5, 1.0, 'Comparison of Emissions With and Without Margins')
```



In addition to studying emissions across industries, we also analyzed the internal relationships between emission factors. This scatter plot compares supply chain emission factors with and without considering marginal effects. Each point represents an industry, with its X value indicating

the emission intensity of the industry without marginal effects in the supply chain, and its Y value representing the emission intensity after including marginal effects.

From the results, most points are close to the diagonal line, indicating that for most industries, marginal effects contribute relatively little to supply chain emissions. This may be because emissions in these industries primarily come from core production stages, while the impact of additional stages such as retail, distribution, and logistics is relatively small.

However, a few industries have points significantly above the diagonal on the Y-axis, meaning their marginal effects contribute more, as their supply chain emission factors increase significantly when marginal effects are included. Meanwhile, some points have high X-axis values, indicating that certain industries already have high supply chain emissions even without considering marginal effects, such as cement manufacturing and livestock farming. In these industries, the impact of marginal effects is relatively small, as their emissions mainly originate from raw material extraction, production, and transportation.

Most points follow an increasing trend, suggesting that as baseline supply chain emissions increase, the contribution of marginal effects to emissions may also increase. However, some industries show a larger increase (farther from the diagonal), indicating that these industries have high carbon emissions in the value-added stages.

The overall conclusion is that the core source of supply chain emissions still lies in the manufacturing stage. Therefore, to reduce greenhouse gas emissions in the future, the key remains in optimizing the supply chain.

##5. Comparison of Emissions of and Without Margins

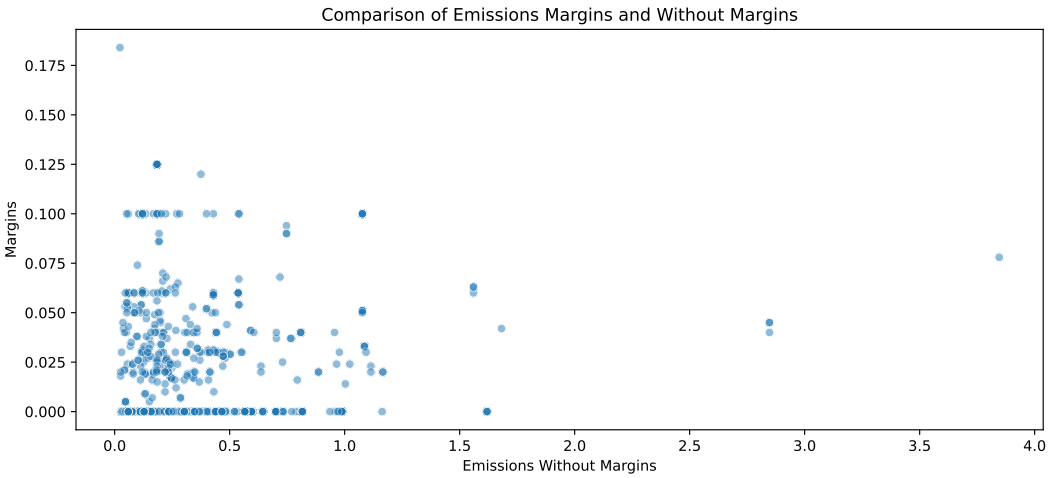
```
## <Figure size 1200x500 with 0 Axes>
```

```
## <Axes: xlabel='Supply.Chain.Emission.Factors.without.Margins', ylabel='Margins.of.Sup
```

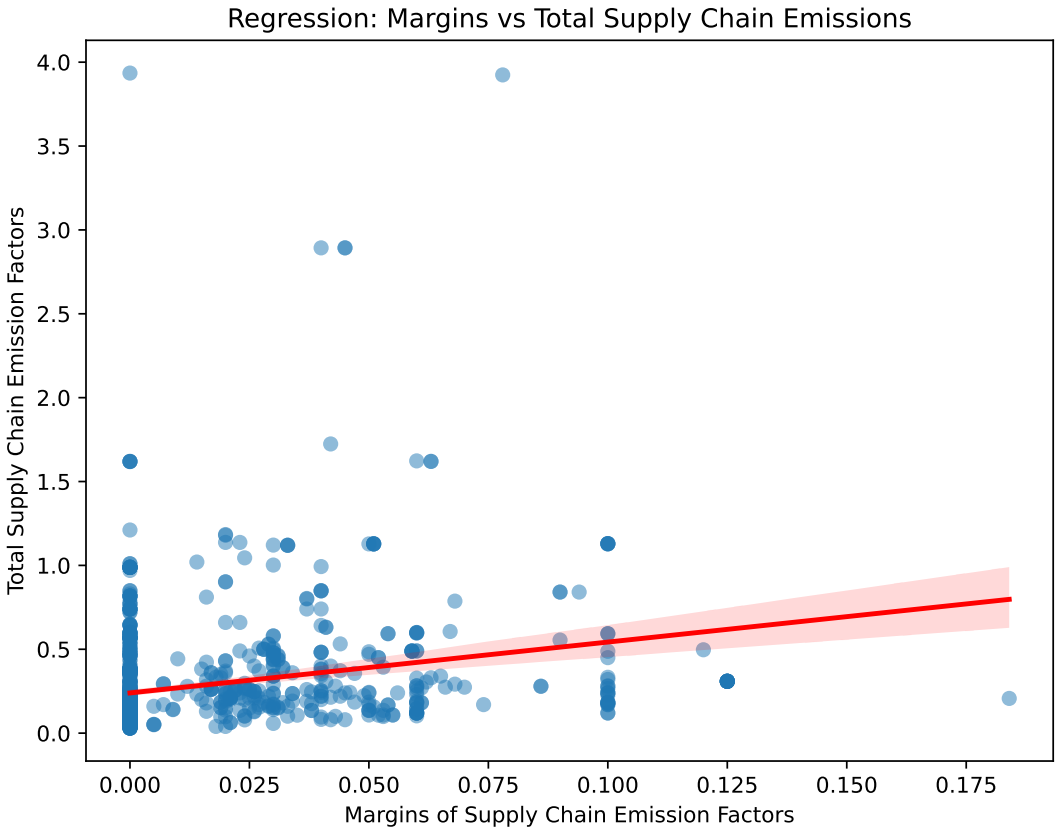
```
## Text(0.5, 0, 'Emissions Without Margins')
```

```
## Text(0, 0.5, 'Margins')
```

```
## Text(0.5, 1.0, 'Comparison of Emissions Margins and Without Margins')
```




```
## <Figure size 800x600 with 0 Axes>

## <Axes: xlabel='Margins.of.Supply.Chain.Emission.Factors', ylabel='Supply.Chain.Emissi

## Text(0.5, 0, 'Margins of Supply Chain Emission Factors')

## Text(0, 0.5, 'Total Supply Chain Emission Factors')

## Text(0.5, 1.0, 'Regression: Margins vs Total Supply Chain Emissions')
```



The fourth chart analyzed the relationship between Emissions with and Without Margins, while this chart illustrates the relationship between the "non-marginal" and "marginal" portions of supply chain emission factors. The purpose of this study is to explore whether there is a linear relationship between the two. However, the results indicate that the correlation between them is not strong.

The scatter plot does not show a clear linear trend, suggesting that high-emission industries (such as cement and steel) do not necessarily have high marginal supply chain emissions. The marginal portion of emissions is more dependent on industry characteristics, such as value-added components, branding, and logistics.

However, a linear regression model indicates that there is a certain degree of correlation between the two. Therefore, carbon emissions research can be conducted separately from the perspectives of production and profitability, while also exploring the intrinsic connection between these two aspects.

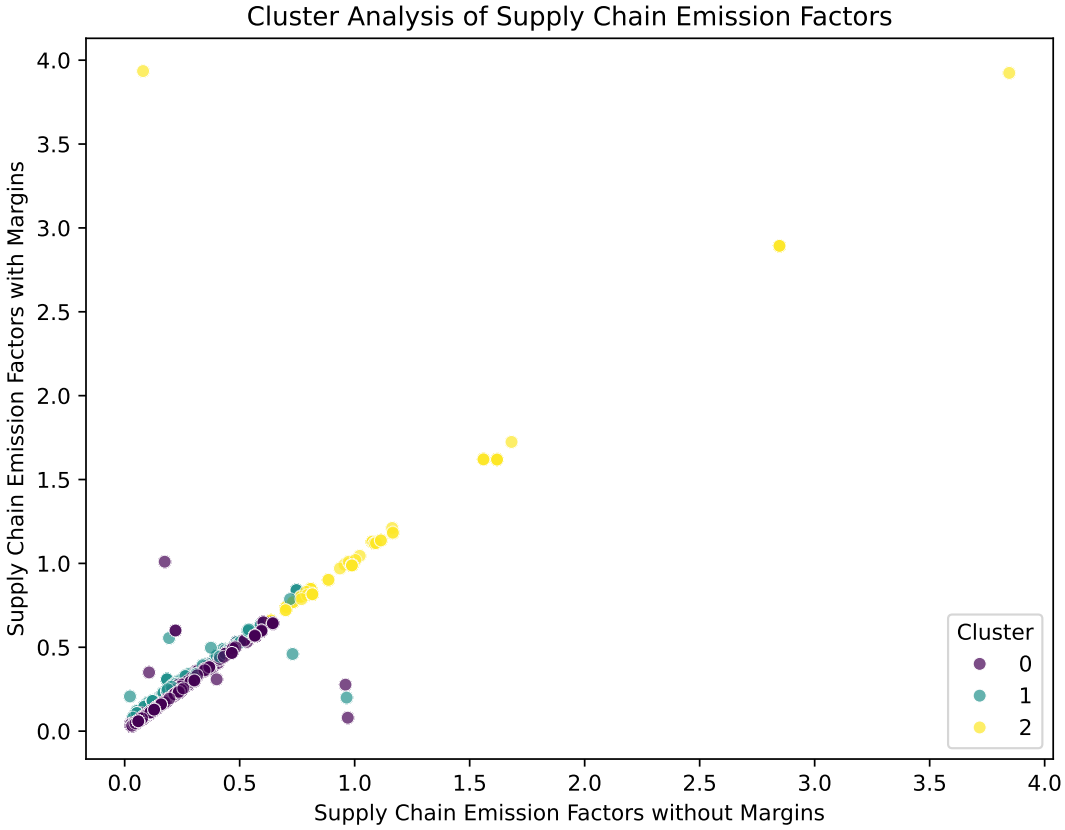
V Feature engineering process and justification

The previous EDA provided an initial exploratory analysis of the data. Now, we will conduct an in-depth analysis using feature engineering. Given the characteristics of this dataset, we have decided

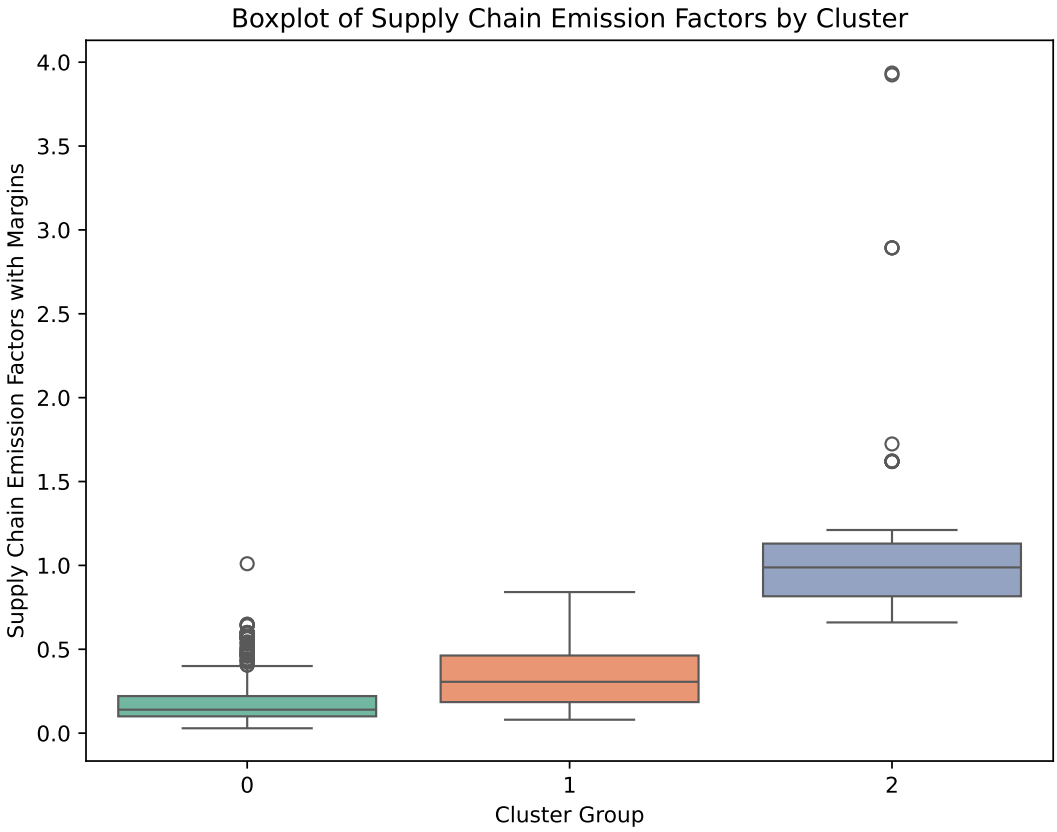
to apply clustering to gain a deeper understanding of supply chain emission factor patterns, industry similarities, and strategies for optimizing emission reduction.

The ultimate goal of clustering analysis is to categorize different industries based on their supply chain emission patterns and identify industries with similar carbon emission characteristics. Therefore, we use three types of carbon emission data from the dataset to perform K-means clustering. The resulting clustering visualization and boxplots for the three categories are shown below.

```
## KMeans(n_clusters=1, n_init=10, random_state=42)
## KMeans(n_clusters=2, n_init=10, random_state=42)
## KMeans(n_clusters=3, n_init=10, random_state=42)
## KMeans(n_clusters=4, n_init=10, random_state=42)
## KMeans(n_clusters=5, n_init=10, random_state=42)
## KMeans(n_clusters=6, n_init=10, random_state=42)
## KMeans(n_clusters=7, n_init=10, random_state=42)
## KMeans(n_init=10, random_state=42)
## KMeans(n_clusters=9, n_init=10, random_state=42)

## <Figure size 800x600 with 0 Axes>

## <Axes: xlabel='Supply.Chain.Emission.Factors.without.Margins', ylabel='Supply.Chain.E

## Text(0.5, 0, 'Supply Chain Emission Factors without Margins')

## Text(0, 0.5, 'Supply Chain Emission Factors with Margins')

## Text(0.5, 1.0, 'Cluster Analysis of Supply Chain Emission Factors')

## <matplotlib.legend.Legend object at 0x000001E1F8850EF0>
```




```
## <Figure size 800x600 with 0 Axes>

## <Axes: xlabel='Cluster', ylabel='Supply.Chain.Emission.Factors.with.Margins'>

## Text(0.5, 0, 'Cluster Group')

## Text(0, 0.5, 'Supply Chain Emission Factors with Margins')

## Text(0.5, 1.0, 'Boxplot of Supply Chain Emission Factors by Cluster')
```



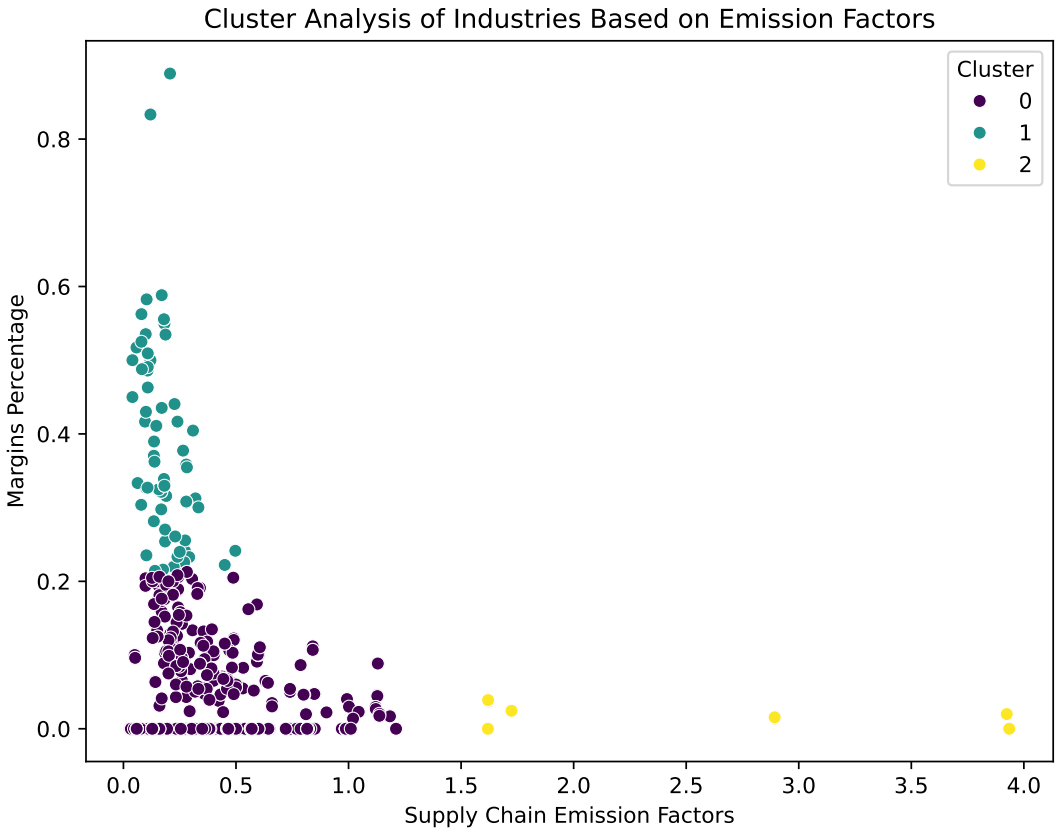
From the scatter plot results, it can be observed that the points align along the diagonal, indicating that for most industries, the supply chain emission factors remain relatively unchanged whether marginal effects are included or not. This suggests that marginal effects have a minimal impact on most industries, and the primary source of supply chain emissions remains the core production process.

Regarding the three categories (0, 1, and 2), the positions of clusters 1 and 2 represent medium-emission and low-emission industries, respectively. However, their data points almost overlap, making them difficult to distinguish. In contrast, cluster 0 represents high-emission industries, which show a significant positional difference from the other two categories, allowing for clear differentiation. Therefore, it can be inferred that these high-emission industries may require stricter emission reduction policies, such as carbon taxes and energy structure adjustments.

The boxplot results further support the findings from the scatter plot. The high-emission industry category exhibits a higher median value, with more dispersed outliers above the upper bound, demonstrating a significant gap compared to the other two categories. The medium-emission industry category has a lower median, with data distributed more consistently, indicating that industries in this category have relatively stable supply chain emissions and moderate overall carbon emissions. Meanwhile, the low-emission industry category has the lowest median value, with a generally lower distribution and the least fluctuation, suggesting a more stable and lower carbon footprint.

Since the initial clustering results did not effectively distinguish between low-emission and medium-emission industries, we applied feature engineering to construct new variables and re-cluster the data, aiming for improved results. Building on insights from the previous exploratory data analysis, we introduced an industry category classification based on industry names, grouping them into four categories: Energy-Intensive, High-Value, Low-Emission, and Other. Additionally, we constructed Margins Percentage, which calculates the proportion of carbon emissions from marginal supply chain activities such as retail, distribution, and logistics. By clustering using Supply Chain Emission Factors and Margins Percentage, we aim to simultaneously differentiate high-emission and low-emission industries as well as production-dominant and supply chain value-added dominant industries. The resulting clustering outcomes are shown below.

```
## <Figure size 800x600 with 0 Axes>

## <Axes: xlabel='Supply.Chain.Emission.Factors.with.Margins', ylabel='Margins_Percentag

## Text(0.5, 0, 'Supply Chain Emission Factors')

## Text(0, 0.5, 'Margins Percentage')

## Text(0.5, 1.0, 'Cluster Analysis of Industries Based on Emission Factors')
```



From the scatter plot, it is evident that the three categories are clearly distinguishable. Category

2 represents high-emission industries, characterized by significantly high emissions but a low marginal contribution rate. Categories 0 and 1 both have relatively low emissions, with category 0 having slightly higher emissions than category 1, but also a higher marginal contribution rate.

This clustering result is superior to the previous approach, which was based solely on emission levels. By incorporating both emission levels and marginal contribution rates, the clustering analysis provides a more comprehensive and intuitive representation of industry differences.

#VI Summary of key findings

In the analysis of this dataset, we conducted exploratory data analysis and feature engineering on industry categories and different types of emissions. Additionally, we applied K-means clustering to further analyze the carbon emission patterns across industries. The study focused on distinguishing emissions across different industry types and examining the relationships between Margins of Supply Chain Emission Factors and Supply Chain Emission Factors without Margins.

After summarizing our findings, we conclude the following key points:

- High-emission industries are mainly concentrated in cement manufacturing and livestock industries (such as cattle ranching, feedlots, and dairy cattle production). These industries exhibit the highest supply chain emission factors, with significantly higher carbon emissions per unit of economic output compared to other industries, reaching approximately 3.0 - 4.0 kg CO_2e/USD.
- The contribution of marginal supply chain emissions is relatively low compared to overall emission factors, indicating that the primary emissions originate from core production processes rather than logistics, retail, and other value-added activities. Industries with high added value (such as apparel, garden equipment, and software) have a relatively higher marginal emission share, suggesting that optimizing supply chains and green logistics could further reduce their carbon footprint.
- There is a positive correlation between supply chain emission factors and those without marginal effects, meaning that for most industries, incorporating marginal effects does not significantly alter the overall emission factor.
- Clustering results indicate that: 1. High-emission industries primarily include cement, steel, and livestock industries, which are energy-intensive sectors with extremely high supply chain emissions but low marginal effects.2. Medium-emission industries include manufacturing and food processing, with moderate emissions and relatively stable supply chain emissions. 3. Low-emission industries consist of consulting, IT, and finance, which have the lowest carbon emissions but relatively higher marginal contributions.

#VII Challenges faced and future recommendations

For the analysis of this dataset, we have identified the following shortcomings and areas for improvement:

- Limited data sources and insufficient numerical content: The dataset is panel data with only three types of emission factors, which may result in an incomplete analysis. To enhance data comprehensiveness, multiple data sources should be integrated, and time series data should be introduced to analyze the dynamic changes in supply chain carbon emissions.

- Significant intra-industry differences, making precise classification difficult: Even after applying feature engineering, some industries remain difficult to categorize accurately. Additionally, the second round of feature construction was somewhat rough and lacked rigor. More features could be introduced, such as industry energy consumption structure, supply chain length, and product lifecycle carbon footprint, to improve classification accuracy. Furthermore, alternative clustering methods (such as DBSCAN or hierarchical clustering) should be explored to better identify subcategories within industries rather than relying solely on K-means clustering.
- Insufficient differentiation in clustering results, particularly between low-emission and medium-emission industries: The current clustering approach relies solely on supply chain emission factors, ignoring economic activities, added value, and supply chain complexity, which may also influence carbon emissions. In the future, constructing a richer set of features or applying supervised learning models (such as Random Forest or XGBoost) could help classify industries more accurately by leveraging known industry labels.

#VIII Members

Anika Kathuria: Found the dataset and performed cleaning processes on the data.

Yixuan Chen: Mess the data, part of data cleaning, EDA and Feature engineering process and justification, write the word type draft.

Nan Xiao: Data cleaning

Obiora Okeke: Data cleaning, Data checking.

Github link: https://github.com/anikakathuria/ADS_Group9