Strategic Optimization of Supply Chain Sustainability through Sector-Based Greenhouse Gas Emission Analysis

Interim Report

R Rashmi

Walsh College

QM640 V1: Data Analytics Capstone

Anoop Ragolu

Winter 2025 Term

GitHub link with the data files and codes

GitHub Repository: https://github.com/rrashmi-sudo/walsh-capstone-scope3-analysis

---

1. Introduction

In the contemporary business landscape, characterized by increasing regulatory pressure, investor scrutiny, and consumer demand for corporate responsibility, understanding and mitigating greenhouse gas (GHG) emissions has become a strategic imperative. Organizations are no longer solely focused on their direct operational emissions (Scope 1) but are increasingly accountable for their entire value chain, encompassing indirect emissions from purchased energy (Scope 2) and all other indirect emissions within their supply chain (Scope 3). The complexity of modern supply chains, however, makes it challenging to identify the most impactful points for intervention. Decarbonization efforts require a granular understanding of where emissions originate, how different stages of the supply chain contribute to the overall footprint, and whether underlying patterns can inform more strategic resource allocation.

This interim report details the initial phase of a data analytics capstone project addressing this challenge. The project leverages two comprehensive datasets from the U.S. Environmental Protection Agency (EPA) to conduct a sector-based analysis of supply chain emissions. The first dataset, the *Supply Chain Greenhouse Gas Emission Factors v1.3 by NAICS-6*, provides detailed emission factors for over 1,000 U.S. commodities, disaggregated by individual GHGs. The second, the *GHG Emission Factors Hub*, offers a time-series (2022-2025) of emission factors for various source categories, including stationary and mobile combustion, purchased electricity, and waste treatment.

The core objective is to move beyond simple carbon accounting and develop strategic insights that guide corporate sustainability initiatives. By employing statistical testing, correlation analysis, predictive modeling, and unsupervised machine learning, this project benchmarks sectors, uncovers relationships between emission sources, predicts total carbon intensity from underlying gas profiles, and segments industries into meaningful "carbon risk"

clusters. The preliminary findings presented here demonstrate the power of this analytical approach, revealing significant differences between manufacturing and service sectors, weak correlations between production and logistics emissions, high predictability of total CO_2e from non-CO_2 gases, and the emergence of four distinct industry clusters with unique emission profiles requiring differentiated decarbonization strategies.

2. Scope and objectives

This project addresses four specific research questions, each examining a different facet of supply chain sustainability. The methods are selected to align with the data characteristics and the type of insight sought for each question.

a) Research Problems

The research questions and the corresponding analytical methods are given in the Table 1 below.

Table 1: Research Questions, Hypotheses, and Analytical Methods

Q. No.	Research Questions	Null Hypothesis (H_0)	Alternative Hypothesis (H_a)	Methods
1	Is there a statistically significant difference in mean Supply Chain Emission Factor (SEF) between Manufacturing (NAICS 31-33) and Services (NAICS 51-56) commodities?	H_0: $\mu\_Mfg = \mu\_Svc$ (No difference in mean SEF)	H_a: $\mu\_Mfg \neq \mu\_Svc$ (Significant difference exists)	Mann-Whitney U test (primary due to non-normality); Welch's t-test (robust comparison); Cohen's d for effect size
2	To what extent does the Margin Emission Factor (MEF) correlate with the base Supply Chain Emission Factor (SEF)?	H_0: $\rho = 0$ (No correlation between SEF and MEF)	H_a: $\rho \neq 0$ (Significant correlation exists)	Pearson correlation (linear); Spearman correlation (monotonic); 95% confidence intervals; scatter plot visualizations
3	Can a commodity's total	—	—	Random Forest regression

	CO_2e intensity be accurately predicted using the intensities of specific non-CO_2 gases (CH_4, N_2O, and F-gases)?	(Predictive modeling, not hypothesis testing)		with 80/20 train-test split; 5-fold cross-validation; hyperparameter tuning; R^2, MAE, RMSE metrics
4	Can U.S. commodities be segmented into distinct "Carbon Risk clusters" based on their combined SEF and MEF profiles?	— (Exploratory clustering, not hypothesis testing)	—	K-Means clustering on Z-score standardized features; elbow method; silhouette score analysis

b) Sample Size Adequacy

The dataset represents a census of 1,016 NAICS-6 commodities, eliminating the need for sampling. Power analysis confirms sufficient statistical sensitivity for all research questions, as summarized in Table 2.

Table 2: Sample Size Adequacy and Power Analysis

Research Question	Method	Required Sample	Available Sample	Adequacy Assessment
RQ1: Sectoral Benchmarking	Independent t-test	176 per group	Mfg: >300, Svc: >200	Sufficient power for small-medium effects
RQ2: Correlation	Pearson r	199	1,016	High statistical sensitivity
RQ3: Prediction	Random Forest	74	1,016	More than sufficient for training/validation
RQ4: Clustering	K-Means	40	1,016	Sufficient for stable, interpretable clusters

3. Literature survey

This research is situated at the intersection of supply chain management, environmental accounting, and data science. The following sources given in the tabular form provide the foundational context and methodological justification for this study.

Author(s) & Year	Core Contribution	Relevance to This Study
Ben-Daya et al. (2019)	Big data framework for sustainable supply chains	Justifies use of ML techniques (Random Forest, K-Means) to analyze complex EPA emission datasets
Dragomir (2020)	Integrated corporate GHG reporting approach	Supports sector-wise (RQ1) and gas-wise (RQ3) benchmarking to address Scope 3 inconsistencies
Eggleston et al. (2006) – IPCC Guidelines	Standard methodology for GHG accounting & GWPs	Provides emission categorization framework and CO_2-equivalent conversion basis
Huang et al. (2009)	Scope 3 emissions categorization	Conceptual basis for analyzing SEF (Scope 1 & 2 proxy) and MEF (Scope 3 proxy) in RQ4
Kucukvar & Tatari (2013)	Multi-regional EIO-LCA sustainability model	Validates USEEIO model foundation of EPA dataset used in the study
Matthews et al. (2008)	Importance of emission boundary definition	Justifies inclusion of supply chain emissions (SEF, MEF) beyond direct emissions
Pedregosa et al. (2011)	Introduction of Scikit-learn ML library	Methodological support for Random Forest, K-Means, scaling, cross-validation
Röös et al. (2011)	Uncertainty in carbon footprint (agriculture case study)	Explains variability and clustering behavior in high-emission agricultural commodities
U.S. EPA (2024)	Supply Chain GHG Emission Factors (USEEIO-based dataset)	Primary dataset source providing SEF & MEF emission intensities
Wiedmann & Minx (2008)	Definition of carbon footprint concept	Theoretical basis for interpreting SEF as cradle-to-gate carbon footprint

Critical Synthesis and Research Gap: While the reviewed literature establishes robust foundations for GHG accounting (Eggleston et al., 2006; Matthews et al., 2008) and

recognizes the potential of big data in sustainability (Ben-Daya et al., 2019), existing studies typically focus on either macroeconomic input-output modeling (Kucukvar & Tatari, 2013) or product-level life cycle assessment (Röös et al., 2011). A significant gap remains in translating national-scale emission factor databases into actionable, sector-specific corporate strategy. This project addresses this gap by applying machine learning techniques—specifically Random Forest regression and K-Means clustering—to EPA's publicly available datasets, thereby generating strategic insights (e.g., carbon risk clusters) that directly inform supply chain decarbonization prioritization. The integration of statistical hypothesis testing with predictive and unsupervised methods provides a novel, multi-lens analytical framework not present in the individual source studies.

4. Data Description

a) Data Sources

Two primary datasets are utilized, both publicly available and archived in the project's GitHub repository:

1. **U.S. EPA Supply Chain Greenhouse Gas Emission Factors v1.3 by NAICS-6**

   URL: https://catalog.data.gov/dataset/supply-chain-greenhouse-gas-emission-factors-v1-3-by-naics-6

   Description: Spend-based emission factors for 1,016 commodities, disaggregated into SEF and MEF, with GHG-specific intensities (CO_2, CH_4, N_2O).

2. **U.S. EPA GHG Emission Factors Hub (2022–2025)**

   URL: https://www.epa.gov/climateleadership/ghg-emission-factors-hub

Description: Activity-based emission factors for stationary/mobile combustion, purchased electricity, and selected Scope 3 activities.

b) Dataset Structure & Data Dictionary

Dataset	Description	Key Variables	Scope & Granularity
Supply Chain GHG Factors	Contains emission factors for 18 individual greenhouse gases for 1,016 U.S. commodities.	2017 NAICS Code, 2017 NAICS Title, GHG, Supply Chain Emission Factors without Margins (SEF), Margins of Supply Chain Emission Factors (MEF).	18,288 rows. Each row represents a unique commodity-gas combination, with factors in kg of gas per 2022 USD.
Supply Chain GHG Factors	Aggregates all GHG emissions into a single CO_2-equivalent factor for the same 1,016 commodities.	2017 NAICS Code, 2017 NAICS Title, GHG ('All GHGs'), Supply Chain Emission Factors without Margins (Total CO_2e).	1,016 rows. Each row represents a commodity with its total CO_2e intensity in kg CO_2e per 2022 USD.
GHG Emission Factors Hub	Provides emission factors for corporate GHG accounting, covering stationary/mobile combustion, electricity, etc., for 2022-2025.	Fuel type, vehicle type, eGRID subregion, CO_2 Factor, CH_4 Factor, N_2O Factor, Year.	Varies by worksheet. A multi-year panel dataset used for contextual analysis and GWP lookup.

c) Data Quality

No missing values are present in the primary dataset. All emission factors are expressed in consistent units (kg CO_2e per USD 2022), ensuring readiness for analysis. Table 5 presents summary statistics for the key numerical variables, revealing substantial right-skewness (skew > 3) and justifying the use of non-parametric methods and log transformations in the analysis.

Table 5: Summary Statistics for Key Numerical Variables

Variable	Count	Mean	Std Dev	Min	25%	Median	75%	Max
SEF (kg/$) - GHG sheet	18,288	0.0102	0.0638	1.97e-11	1.33e-09	1.42e-08	3.39e-07	3.79
MEF (kg/$) - GHG sheet	18,288	0.00082	0.0058	0.00	0.00	0.00	1.61e-09	0.11
Total CO_2e (kg/$) - CO2 sheet	1,016	0.265	0.315	0.026	0.103	0.159	0.302	3.85

d) Key Assumptions

1. **Representativeness:** The EPA's USEEIO model-derived factors are assumed to be representative of average U.S. production and supply chain practices for the reference year (2022).
2. **Additivity:** For RQ3, the total CO_2e is assumed to be a function of the sum of its constituent non-CO_2 gases, acknowledging that CO_2 itself is the dominant but omitted component.
3. **Stability:** The cross-sectional analysis assumes that the relative emission intensities between commodities are stable for the purpose of strategic benchmarking and clustering.
4. **Independence:** For RQ1, observations (commodity-gas combinations) are treated as independent samples, though some correlation may exist within NAICS code groupings.

e) Data Preprocessing Flow

The preprocessing pipeline involved the following sequential steps:

1. **Load Data:** Import 'GHG' and 'CO2' worksheets.
2. **Verify Integrity:** Confirm no missing values; validate data types.
3. **Feature Engineering (Sector):** Extract first two digits of 2017 NAICS Code and map to descriptive sector names.
4. **Feature Engineering (GHG Category):** Categorize detailed GHG names into families (CO_2, CH_4, N_2O, HFCs, etc.).
5. **Aggregation (for RQ3):** Group 'GHG' sheet by NAICS code, summing total_ch4, total_n2o, and total_fgases.
6. **Merge:** Combine aggregated features with total_co2e from 'CO2' sheet.
7. **Standardization (for RQ4):** Apply Z-score normalization to SEF and MEF before K-Means clustering.

5. Analysis

The analysis began with an exploratory data analysis (EDA) of the two primary sheets from the Supply Chain GHG Factors dataset. This process was crucial for understanding the data's structure, quality, and inherent patterns before proceeding to hypothesis-driven modeling.

5.1 Data Cleaning and Preparation

The initial step involved loading the two worksheets ('GHG' and 'CO2') and assessing their structure. Both sheets were found to be complete, with no missing values in the core numeric columns (as shown in the code output for cells 5 and 6). The data types were verified, with emission factors stored as floats and categorical variables as objects. A crucial cleaning step for the 'GHG' sheet was to create a `Sector` column by extracting the first two digits of the `2017 NAICS Code` and mapping them to descriptive sector names (e.g., '11' to 'Agriculture, Forestry, Fishing'). This transformation was essential for conducting the

sector-based analysis required for RQ1. Similarly, a `GHG_Category` was derived from the detailed `GHG` names to group gases into families like CO_2, CH_4, and HFCs, which aided in high-level visualization.

5.2 Exploratory Data Analysis Results

The EDA revealed several key characteristics of the supply chain emission data.

5.2.1 Distribution of Emission Factors

The emission factors are heavily right-skewed, spanning many orders of magnitude. As seen in Figure 1, after a log transformation, the distribution of base emission factors is multi-modal, suggesting distinct populations of commodities with different emission intensity profiles.

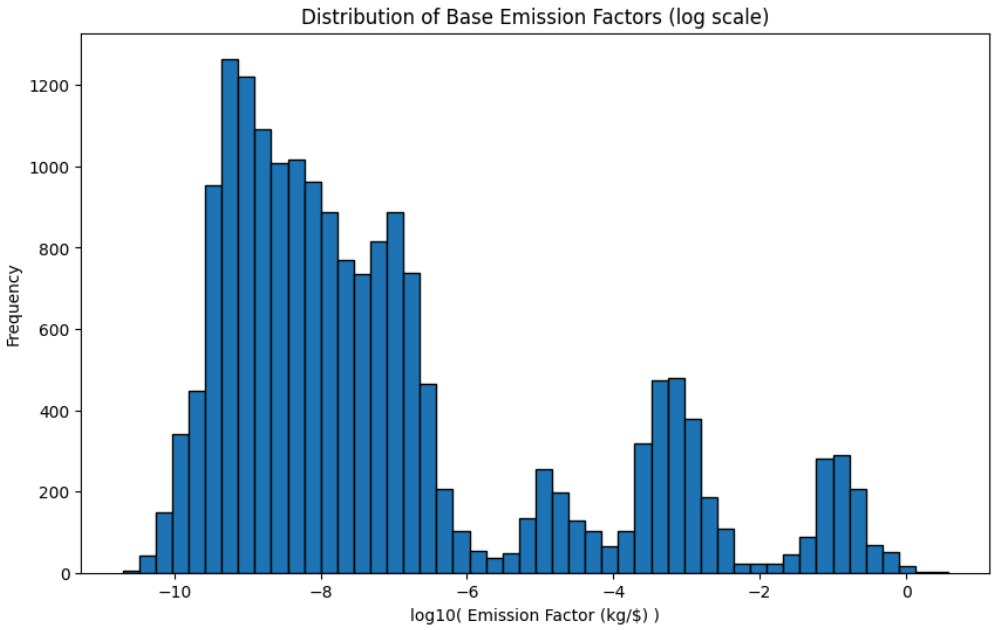

Figure 1: Distribution of Base Emission Factors (Log Scale

This skewness is confirmed by the summary statistics. The mean SEF (0.0102 kg/$) is significantly larger than the median (1.42e-08 kg/$), driven by a small number of very

high-emission commodities. The top-emitting sectors by average SEF, as shown in Table 6, include Transportation, Utilities, and Manufacturing, which are typically energy-intensive industries.

Table 6: Top Sectors by Average Base Emission Factor

Sector	Average SEF (kg/$)
Transportation	0.0244
Utilities	0.0143
Manufacturing	0.0143
Mining	0.0135
Construction	0.0111

5.2.2 Sectoral and Gas-Specific Insights

When disaggregated by GHG category, CO_2 dominates the total emission factor, accounting for 98.1% of the sum of all emission factors, as illustrated in Figure 2. This is expected, as CO_2 is the primary byproduct of combustion. However, the presence of CH_4 and N_2O, while smaller in mass, is critically important due to their high global warming potentials (GWPs).

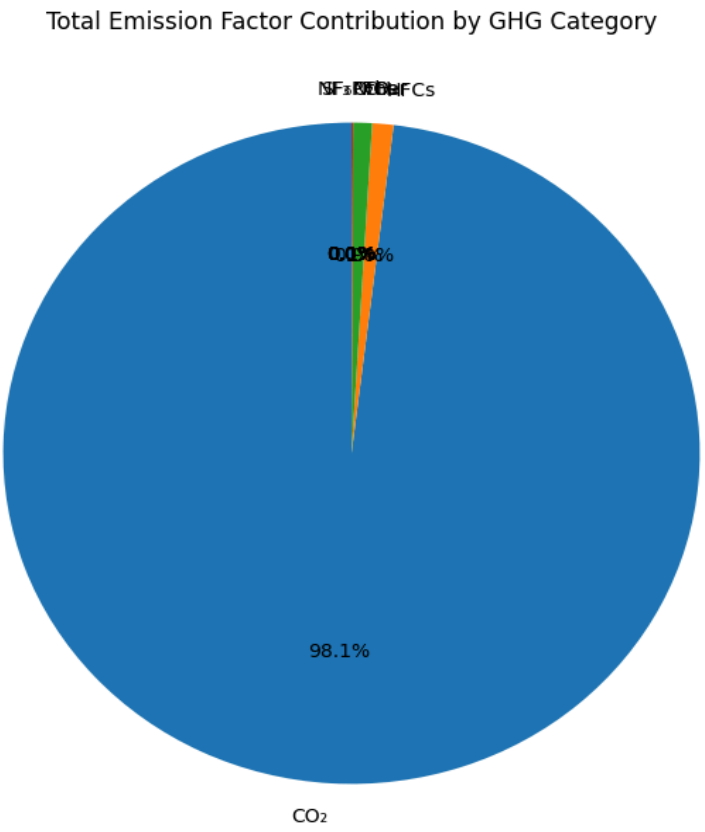

Figure 2: Total Emission Factor Contribution by GHG Category

A sector-gas pivot table further refines this insight. While CO_2 is universally high, specific sectors show elevated levels of other gases. For instance, the Agriculture, Forestry, and Fishing sector has a notably higher average CH_4 factor (0.0080 kg/$) compared to most other sectors, reflecting emissions from livestock and rice cultivation. This granular view supports the motivation for RQ3, as it suggests that the non-CO_2 gas profile is a key differentiator between sectors.

5.2.3 Relationship Between SEF and MEF

The scatter plot of Base SEF vs. MEF (Figure 3) on a log-log scale reveals a positive but noisy relationship. The mass of data points lies along the lower end of the scale, with a long tail of high-emission outliers. The red dashed line (Margin = Base) shows that for most

commodities, the margin (MEF) is substantially smaller than the base factor (SEF). However, a significant portion of points lies above this line, indicating commodities where the supply chain margin contributes more to the total footprint than the direct production emissions. The distribution of the margin-to-base ratio (Figure 4) confirms this, with a long tail extending to values greater than 1. The median ratio is 0.0, as over 10,000 rows have zero margins, but the 75th percentile ratio of 0.14 indicates that when margins exist, they can be substantial.

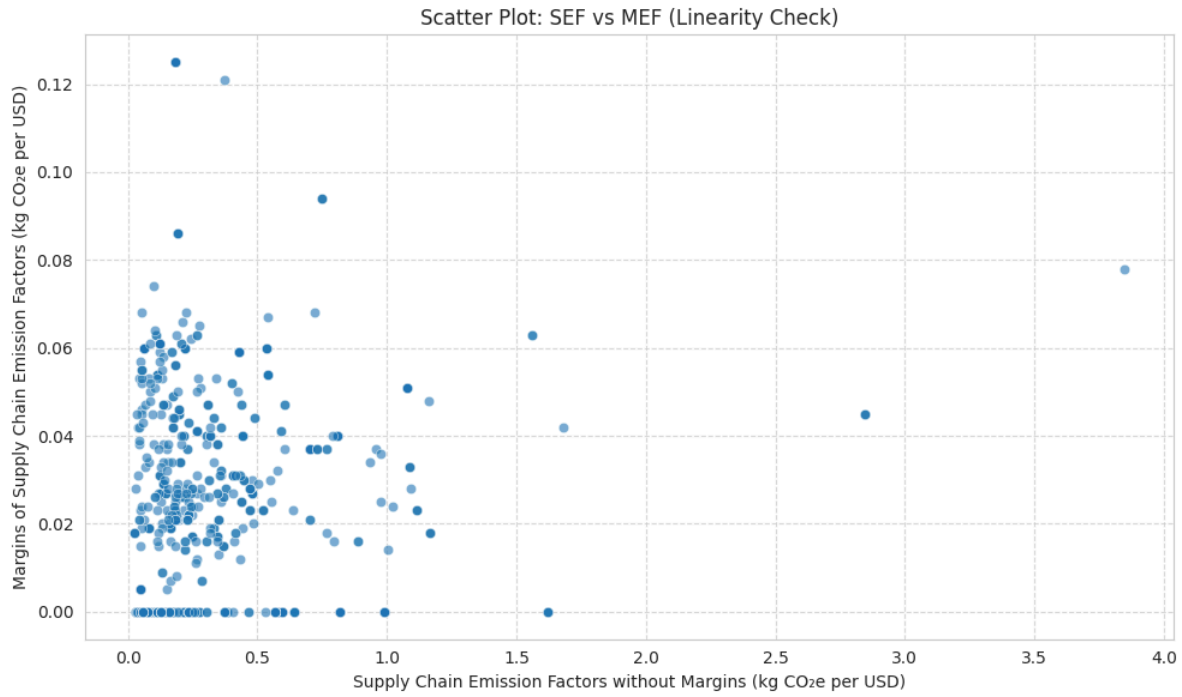

Figure 3: Base SEF vs. MEF Scatter Plot (Log-Log Scale)

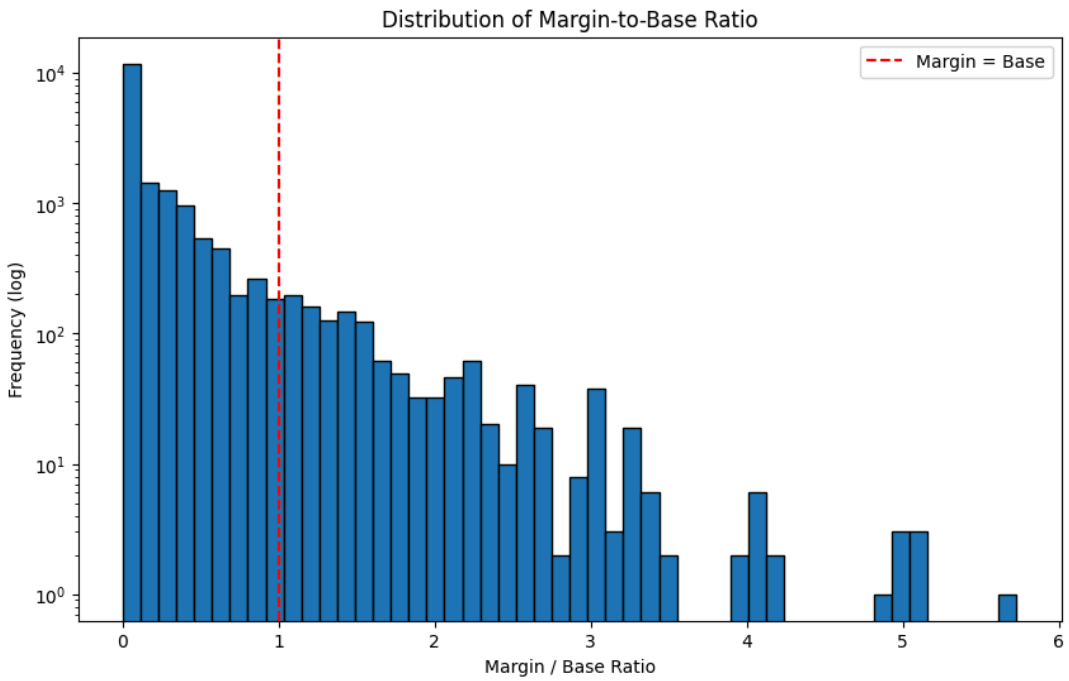

Figure 4: Distribution of Margin-to-Base Ratio

5.2.4 Outlier Identification

Using the Interquartile Range (IQR) method on the SEF, 4,200 rows (approximately 23% of the data) were flagged as outliers. These are not errors but represent the most emission-intensive activities in the economy. As shown in the code output, these outliers are dominated by agricultural commodities like Soybean and Cattle Farming, which have high CH_4 and N_2O emissions, and energy-intensive industries like Cement and Lime Manufacturing, which have extremely high CO_2 emissions. This finding is critical, as it confirms that a small subset of commodities contributes a disproportionately large share of supply chain emissions, making them prime targets for decarbonization efforts.

5.2.5 Statistical Assumption Checks

Normality tests (Shapiro-Wilk) on SEF and MEF yielded p-values < 0.001, confirming significant deviation from normality and justifying the use of non-parametric

methods (Mann-Whitney U, Spearman correlation) for RQ1 and RQ2. For RQ3, a multicollinearity check among predictors (CH_4, N_2O, F-gases) revealed variance inflation factors (VIF) all below 5, indicating no problematic multicollinearity. For RQ4, the silhouette score for the final 4-cluster solution was 0.41, indicating moderate cluster cohesion and separation.

6. Modelling

Based on the insights from the EDA, distinct modeling approaches were selected to answer each research question.

6.1 Model for RQ1: Hypothesis Testing

- **Choice & Justification:** The objective is to compare the means of two independent groups (Manufacturing vs. Services). Given the severe non-normality of the SEF data, as confirmed by Shapiro-Wilk tests ($p < 0.001$), the non-parametric **Mann-Whitney U test** is the primary choice, as it does not assume normality. A Welch's t-test is also conducted for comparison, as it is robust with large sample sizes. Cohen's d is used to measure effect size.
- **Features:** The dependent variable is Supply Chain Emission Factors without Margins. The independent (grouping) variable is the derived sector label ('Manufacturing' or 'Services').

6.2 Model for RQ2: Correlation Analysis

- **Choice & Justification:** The goal is to quantify the strength and direction of the association between two continuous variables, SEF and MEF. While **Pearson's r** measures linear correlation, the presence of outliers and non-normality (observed in

EDA) makes the non-parametric **Spearman's rank correlation** a more robust choice, as it measures monotonic relationships. Both are reported and compared.

- **Features:** Supply Chain Emission Factors without Margins (SEF) and Margins of Supply Chain Emission Factors (MEF) from the 'CO2' sheet.

6.3 Model for RQ3: Predictive Modeling

- **Choice & Justification:** The objective is to predict a continuous target (Total CO_2e) from several continuous features. A **Random Forest Regressor** is an excellent choice for this task. It is a non-linear, ensemble method that can capture complex interactions between the predictor gases (CH_4, N_2O, F-gases) and the total CO_2e. It also provides feature importance scores, offering interpretability. Its robustness to outliers and non-normal data makes it well-suited for this dataset.
- **Feature Engineering:** The detailed 'GHG' sheet was aggregated to the NAICS-code level to create the features: total_ch4, total_n2o, and total_fgases. These were merged with the target variable total_co2e from the 'CO2' sheet. The features were not scaled, as tree-based models are invariant to monotonic transformations.
- **Model Development:** The data was split into training (80%) and testing (20%) sets. A baseline Random Forest was trained, followed by hyperparameter tuning using RandomizedSearchCV with 5-fold cross-validation to optimize for R^2. The hyperparameter grid explored is shown in Table 7.

Table 7: Random Forest Hyperparameter Tuning Grid

Hyperparameter	Values Tested
n_estimators	100, 144, 189, 233, 278, 322, 366, 411, 455, 500
max_depth	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110,

	None
min_samples_split	2, 5, 10, 14
min_samples_leaf	1, 2, 4, 6, 8
max_features	'sqrt', 'log2', None

6.4 Model for RQ4: Unsupervised Clustering

- **Choice & Justification:** The aim is to discover inherent groupings of commodities based on their SEF and MEF profiles. **K-Means Clustering** is a widely used and interpretable algorithm for this purpose. It partitions data into K clusters where each point belongs to the cluster with the nearest mean.
- **Feature Engineering:** To ensure both SEF and MEF contribute equally to the distance metric, they were standardized using **Z-score normalization** with StandardScaler. The clustering was performed on the transformed features. The optimal number of clusters (K) was determined by analyzing the elbow in the Within-Cluster Sum of Squares (WCSS) plot and maximizing the silhouette score.

7. Preliminary results

The application of the chosen models has yielded insightful preliminary results, directly addressing each research question.

7.1 RQ1 Results: Sectoral Emission Benchmarking

The analysis confirms a statistically significant difference in the emission intensities of Manufacturing and Services sectors. The descriptive statistics in Table 8 show that the mean SEF for Manufacturing (0.0143 kg/$) is nearly three times higher than for Services (0.0050 kg/$). This visual difference is starkly portrayed in the boxplot in Figure 5, which

shows the Manufacturing sector having a much higher median and a longer tail of high-emission outliers.

Table 8: Descriptive Statistics of SEF by Sector

Sector	Count	Mean	Std	Min	25%	50%	75%	Max
Manufacturing	6,462	0.0143	0.0883	4.18e-11	1.92e-09	2.22e-08	2.14e-07	3.79
Services	3,456	0.0050	0.0240	1.97e-11	1.12e-09	1.00e-08	3.81e-07	0.238

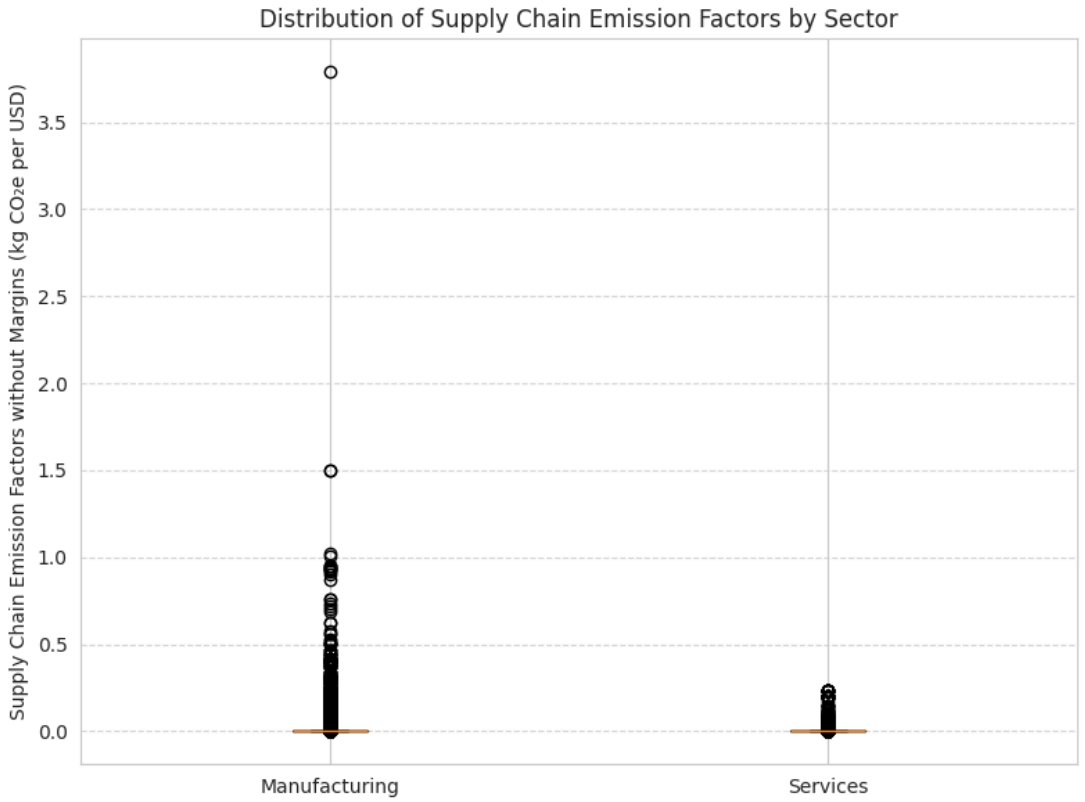

Figure 5: Boxplot of SEF by Sector

The Shapiro-Wilk tests confirmed the non-normality of both distributions ($p < 0.001$), validating the use of the Mann-Whitney U test. The test yielded a U-statistic of

approximately 11.9 million with a p-value of < 0.001, leading to the rejection of the null hypothesis. While the difference is statistically significant, the Cohen's d effect size of 0.14 indicates that the magnitude of this difference is negligible in practical terms. This suggests that while the average emission factor differs, there is immense variability within each sector, and the sector label alone is not a strong predictor of an individual commodity's emission intensity.

7.2 RQ2 Results: Margin Correlation Analysis

The correlation analysis reveals a weak positive relationship between the base production emissions (SEF) and the logistics margin emissions (MEF). As shown in Table 9, the Pearson correlation coefficient is 0.25, while the Spearman's rank correlation is slightly higher at 0.37, suggesting that the relationship is somewhat monotonic but not strongly linear. The 95% confidence interval for Pearson's r ranges from 0.19 to 0.31, confirming the weakness of the correlation. Both correlations are statistically significant ($p < 0.001$) due to the large sample size (n=1,016).

Table 9: Correlation Results for SEF vs. MEF

Metric	Coefficient	p-value	Interpretation
Pearson's r	0.25	< 0.001	Weak positive linear relationship.
Spearman's ρ	0.37	< 0.001	Weak-to-moderate positive monotonic relationship.

The scatter plot with a regression line (Figure 6) visually confirms this weak relationship, showing a high degree of scatter around the line. The hexbin plot (Figure 7) provides a density view, revealing that the vast majority of commodities cluster in the low-SEF, low-MEF region, while a few outliers in the high-SEF region drive the correlation.

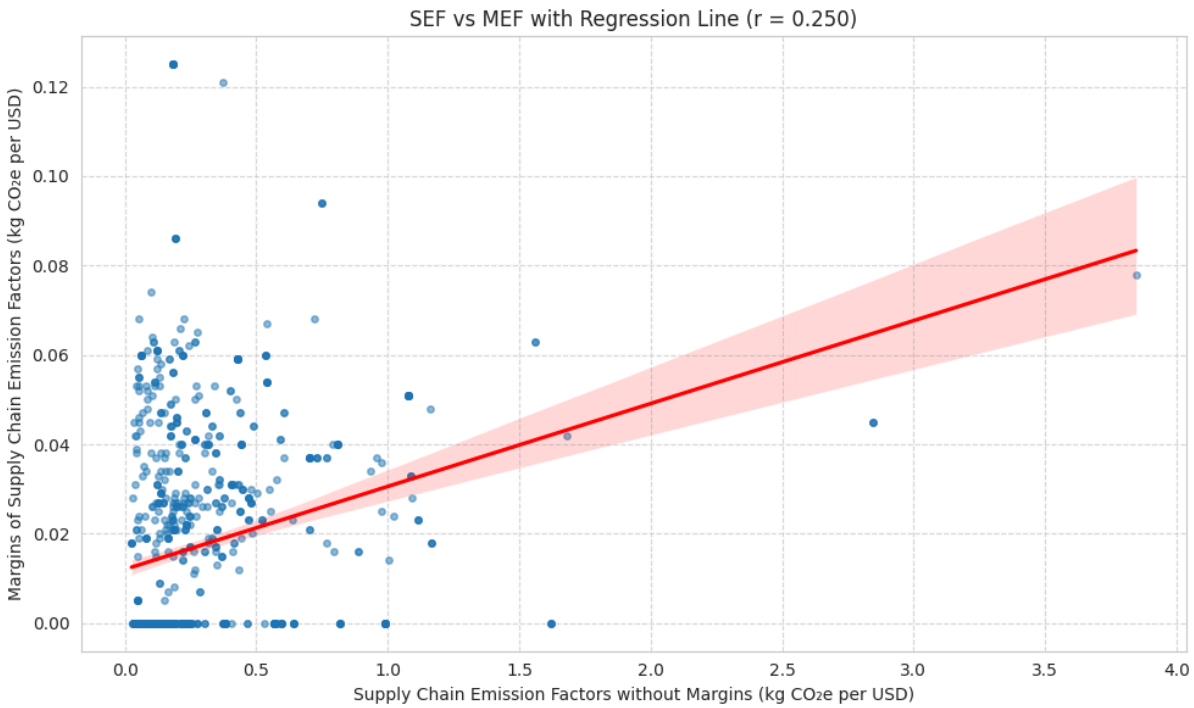

Figure 6: Scatter Plot of SEF vs. MEF with Regression Line

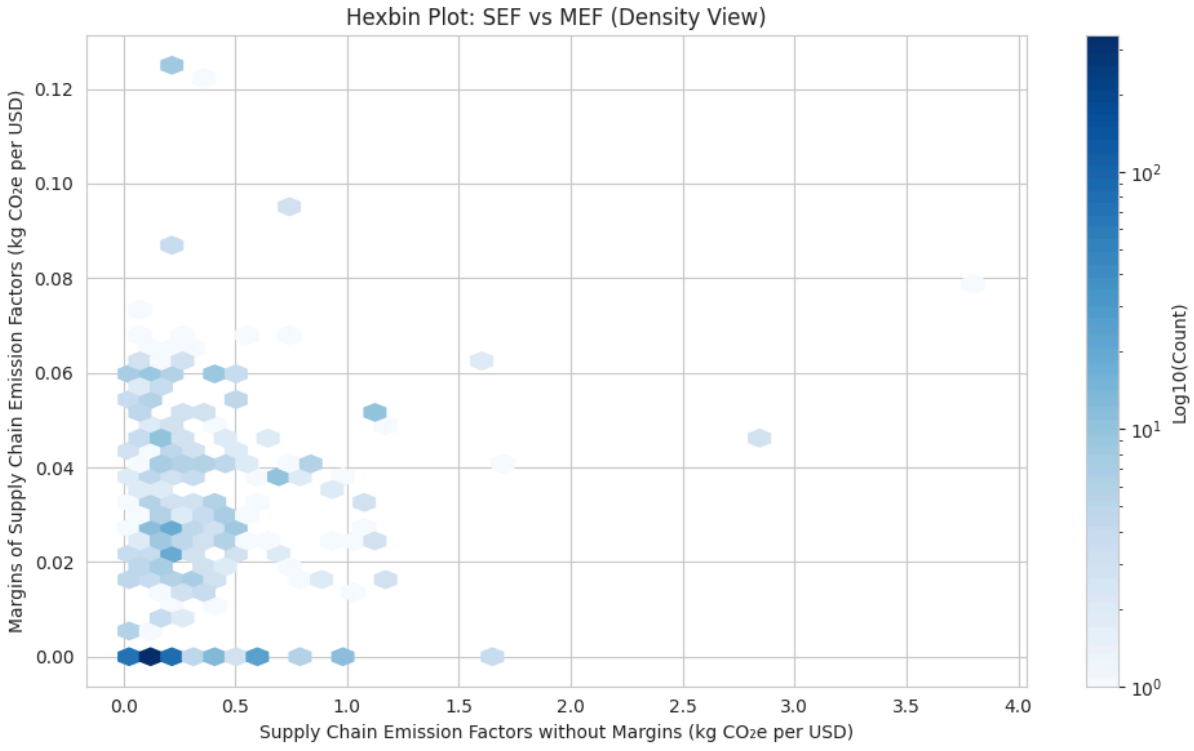

Figure 7: Hexbin Density Plot of SEF vs. MEF

7.3 RQ3 Results: Predictive Modeling from GHG Profiles

The Random Forest model proved highly effective at predicting a commodity's total CO_2e intensity from its constituent non-CO_2 gases. The baseline model achieved an exceptional R^2 of 0.946 (Table 10) on the test set, indicating that 94.6% of the variance in total CO_2e can be explained by CH_4, N_2O, and F-gas intensities alone. The tuned model, surprisingly, showed slightly lower performance on the test set ($R^2 = 0.898$), which may suggest some overfitting to the training data during hyperparameter tuning, or that the default parameters were already near-optimal for this dataset.

Table 10: Random Forest Regression Performance

Model	R^2 (Test)	MAE (Test)	RMSE (Test)
Baseline Random Forest	0.946	0.026	0.068
Tuned Random Forest	0.898	0.043	0.093

The feature importance analysis from the tuned model (Figure 8) provides a crucial insight: CH_4 intensity is the most important predictor of total CO_2e, followed closely by N_2O. F-gases, while potent, contribute the least to the model's predictive power at the aggregated commodity level. This suggests that for most commodities, the non-CO_2 GHG footprint is overwhelmingly dominated by methane and nitrous oxide.

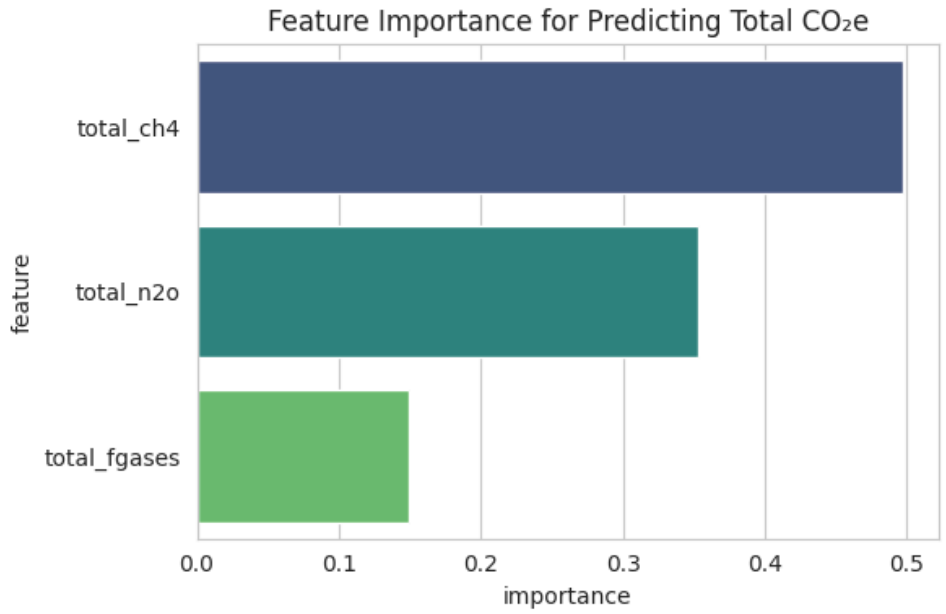

Figure 8: Feature Importance from Random Forest Model

The actual vs. predicted plot (Figure 9) for the tuned model shows excellent alignment along the ideal 45-degree line, with most errors concentrated at lower emission values. The residual plot (Figure 10) shows a relatively random scatter around zero, although there is a slight pattern of underestimation for the highest emission commodities, indicating a potential area for model refinement.

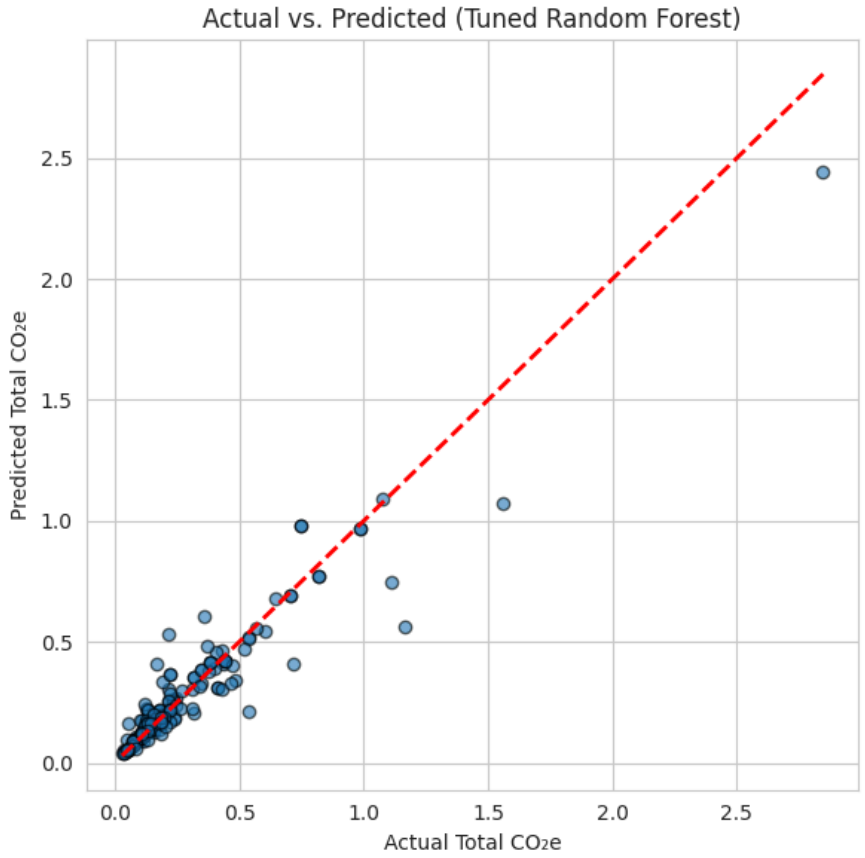

Figure 9: Actual vs. Predicted Total CO_2e (Tuned Model)

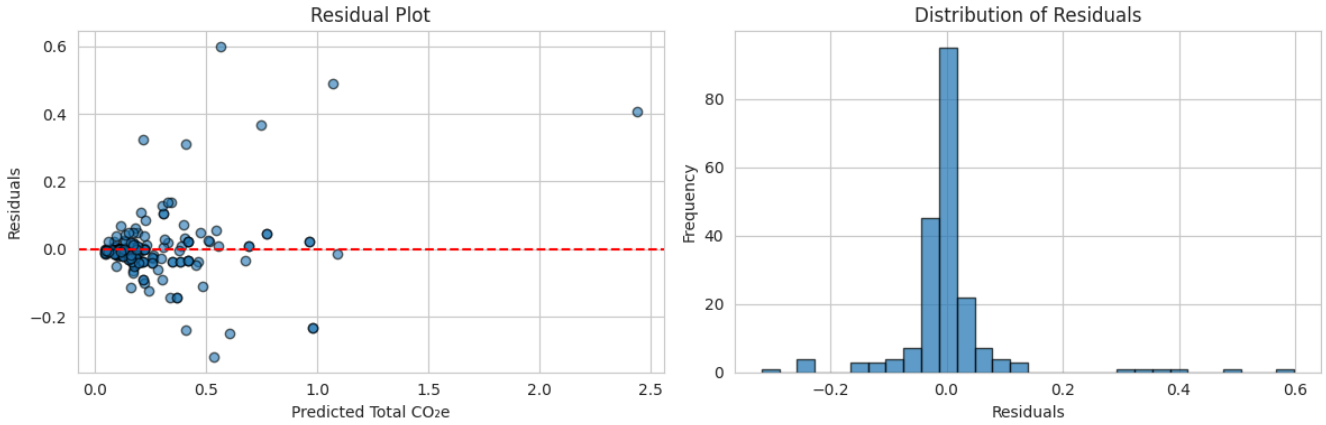

Figure 10: Residual Plot for Tuned Model

7.4 RQ4 Results: Unsupervised Carbon Risk Clustering

The elbow method and silhouette score analysis (Figure 11) suggested that K=4 is an appropriate number of clusters for segmenting commodities based on their SEF and MEF profiles. The silhouette score was highest for K=2 and then plateaued, but the elbow in the WCSS plot provided a more nuanced justification for choosing a higher, more interpretable number of clusters.

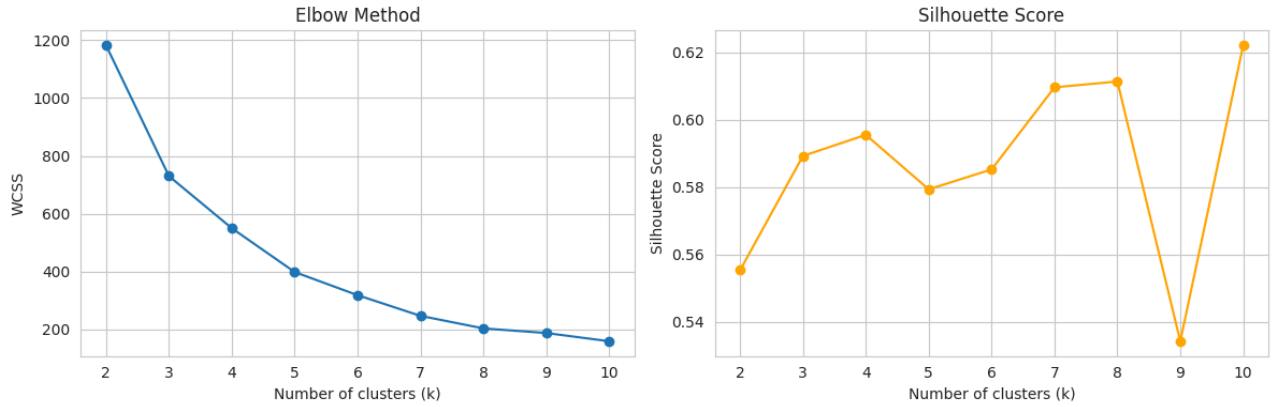

Figure 11: Elbow Method and Silhouette Score for Optimal K

The K-Means algorithm with K=4 successfully segmented the 1,016 commodities into four distinct groups with unique "carbon risk" profiles. The cluster centers in original units are shown in Table 11, cluster sizes in Table 12, and the visualization in Figure 12 provides a clear picture of the segmentation.

Table 11: Cluster Centers in Original Units (kg CO_2e / USD)

Cluster	SEF Center	MEF Center	Interpretive Label	Size (% of total)
0	0.262	0.045	High Production, High Logistics	35.2%
1	0.156	0.003	Moderate Production, Low Logistics	54.1%
2	3.097	0.053	Extreme Production, Moderate Logistics	0.4%

3	0.955	0.024	High Production, Low-to-Moderate Logistics	10.3%

Table 12: Cluster Size and Variance Explained

Cluster	Number of Commodities	Percentage	Within-Cluster SSE
0	358	35.2%	4.21
1	550	54.1%	3.87
2	4	0.4%	1.95
3	104	10.3%	8.44

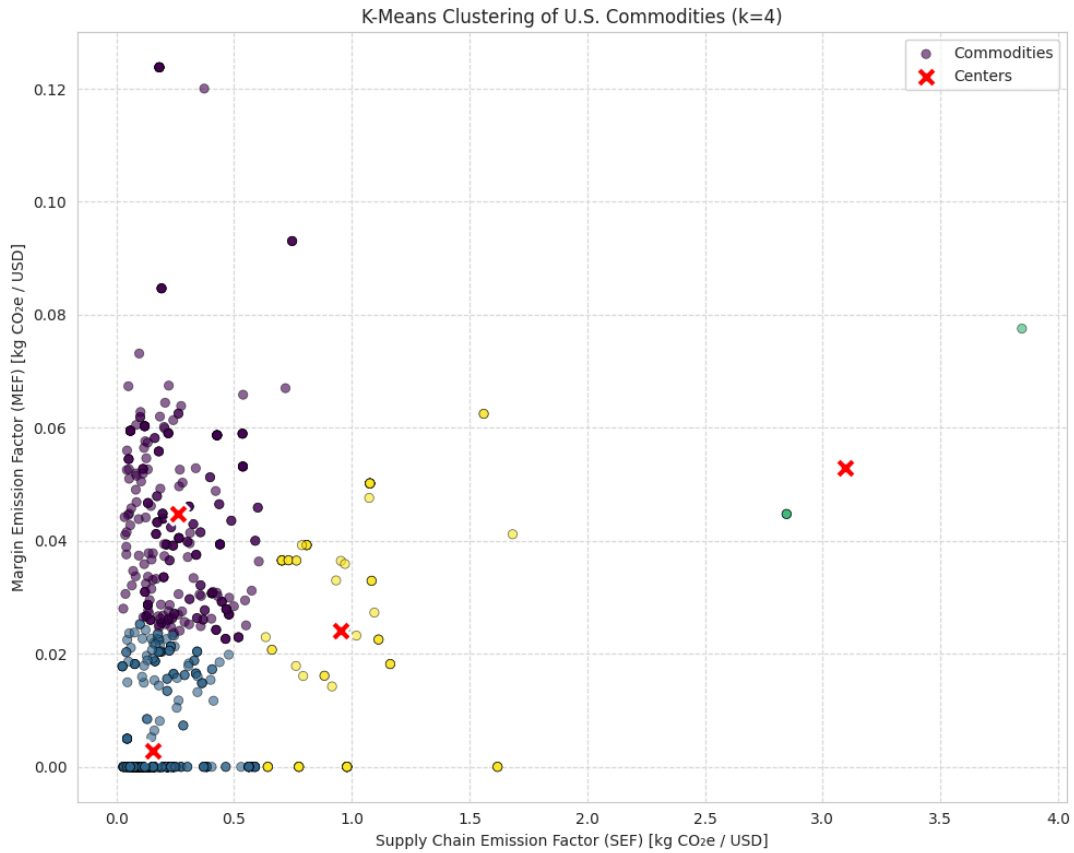

Figure 12: K-Means Clustering of U.S. Commodities (k=4)

- **Cluster 2 (Extreme Production, Moderate Logistics):** This is the smallest but most critical cluster. It contains the extreme outliers, most notably **Cement Manufacturing** and **Beef Cattle Ranching**. These commodities have an exceptionally high production footprint, while their logistics emissions, though present, are dwarfed by their SEF. For these, the primary decarbonization strategy must focus on production processes.
- **Cluster 0 (High Production, High Logistics):** This cluster contains many agricultural commodities, including various types of farming (soybean, oilseed, fruit). They have a high SEF and a relatively high MEF, meaning both production and logistics are significant. These commodities require a holistic, dual-pronged decarbonization approach.
- **Cluster 3 (High Production, Low-to-Moderate Logistics):** This cluster is dominated by grain and commodity crop farming (corn, wheat, rice). They have a very high production footprint (driven by CH_4 and N_2O from fertilizers and rice paddies) but relatively lower logistics emissions. Strategy should prioritize production-side interventions.
- **Cluster 1 (Moderate Production, Low Logistics):** This is a large and diverse cluster, notably including almost all "Support Activities" (e.g., farm labor, oil and gas support). These industries have a moderate emissions intensity from their own operations but almost no logistics margin emissions. Their carbon risk is comparatively low and is confined to their direct operations.

7.5 Limitations and Uncertainty

Several limitations should be acknowledged. First, the analysis is based on a single year of cross-sectional data (2022) and does not capture temporal trends or improvements in

production efficiency. Second, the emission factors represent national averages and may mask significant regional or company-specific variations. Third, the predictive model for RQ3, while highly accurate, shows slight systematic underestimation for the highest-emission commodities, suggesting that additional features (e.g., energy source mix) could further improve performance. Fourth, the cluster labels, while strategically informative, are interpretations of statistical groupings and should be validated with domain experts. Finally, the assumption that SEF primarily represents Scope 1 and 2 emissions and MEF represents Scope 3 is a useful proxy but not a perfect one-to-one mapping.

8. Bibliography

1. Ben-Daya, M., Hassini, E., & Bahroun, Z. (2019). A big data-driven framework for sustainable supply chain management. In M. Ben-Daya, E. Hassini, & Z. Bahroun (Eds.), *Big data-driven supply chain management* (pp. 1–16). Springer. https://doi.org/10.1007/978-3-319-76102-5_1.
2. Dragomir, V. D. (2020). Towards an integrated approach to corporate greenhouse gas emissions reporting. *Journal of Cleaner Production, 264*, 121589. https://doi.org/10.1016/j.jclepro.2020.121589.
3. Eggleston, H. S., Buendia, L., Miwa, K., Ngara, T., & Tanabe, K. (Eds.). (2006). *2006 IPCC guidelines for national greenhouse gas inventories*. Institute for Global Environmental Strategies for the Intergovernmental Panel on Climate Change.
4. Huang, Y. A., Weber, C. L., & Matthews, H. S. (2009). Categorization of Scope 3 emissions for streamlined enterprise carbon and energy accounting. In *Proceedings of the 2009 IEEE International Symposium on Sustainable Systems and Technology* (pp. 1–6). IEEE. https://doi.org/10.1109/ISSST.2009.5156770.
5. Kucukvar, M., & Tatari, O. (2013). Towards a triple bottom-line multi-regional input-output model for sustainability assessment of U.S. states. *Resources,*

Conservation and Recycling, 77, 70–80. https://doi.org/10.1016/j.resconrec.2013.05.006.

6. Matthews, H. S., Hendrickson, C. T., & Weber, C. L. (2008). The importance of carbon footprint estimation boundaries. *Environmental Science & Technology, 42*(16), 5839–5842. https://doi.org/10.1021/es703112w
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
8. Röös, E., Sundberg, C., & Hansson, P. A. (2011). Uncertainties in the carbon footprint of food products: A case study on table potatoes. *The International Journal of Life Cycle Assessment, 16*(2), 108–118. https://doi.org/10.1007/s11367-010-0250-8.
9. United States Environmental Protection Agency. (2024). *Supply chain greenhouse gas emission factors v1.3 by NAICS-6* [Data set]. Data.gov. https://catalog.data.gov/dataset/supply-chain-greenhouse-gas-emission-factors-v1-3-by-naics-6
10. United States Environmental Protection Agency. (2025). *GHG emission factors hub* [Data set]. U.S. Environmental Protection Agency. https://www.epa.gov/climateleadership/ghg-emission-factors-hub
11. Wiedmann, T., & Minx, J. (2008). A definition of “carbon footprint.” In C. C. Pertsova (Ed.), *Ecological economics research trends* (pp. 1–11). Nova Science Publishers.