U.S. Supply Chain's Carbon Footprint: A Data-Driven Approach

Monika Chavan (23288066)*

Friedrich-Alexander-Universität Erlangen-Nürnberg

November 28, 2024

Abstract

This project evaluates the GHG emissions footprints across supply chain sectors, including manufacturing, transportation, and retailing in the United States. Using an automated ETL pipeline, we can significantly streamline the integration of these multiple datasets with robust data quality to explore the detailed trends in emission. Sector-specific differences in carbon outputs are emphasized, along with relevant insights into targeted de-carbonization strategies. Advanced cleaning and transformation techniques made the dataset more usable, while exploratory data analysis highlighted key patterns and intervention opportunities. This work represents an important step toward enabling industries to adopt sustainable practices and reduce their environmental footprint in support of national climate goals. The ambitious future iterations will include real-time emissions tracking and sector-specific normalization to refine insights and expand applicability.

1 Question

How do greenhouse gas emissions footprint vary across different supply chain sectors in the America?

With the increasing focus on sustainability within modern business practices, it is a critical question for understanding the emission across different sectors. Identifying the trend of carbon outputs and disparities that exist will lead to providing actionable insights for industries so they can work on reducing the effects that their businesses have on the environment. The investigation targets sector-specific challenges and opportunities, laying down a data-based groundwork to develop de-carbonization strategies that align with climate goals.

2 Data Sources

To understand greenhouse gas (GHG) emissions across America's supply chains, this study uses four carefully selected datasets, spanning over three decades (1990–2022), to uncover emission trends and identify key contributors.

2.1 Supply Chain Greenhouse Gas Emission Factors v1.3 by NAICS-6

Data Type: CSV
Source: [1]
This dataset categorizes emission factors by NAICS-6 codes, offering insights into sector-specific GHG emissions, essential for comparing emissions across industries.

2.2 Supply Chain Greenhouse Gas Emission Factors for US Industries and Commodities

Data Type: CSV
Source: [2]
This dataset provides a commodity-level view of emissions, helping assess the environmental impact of different materials and products within supply chains.

2.3 Transportation-Related Greenhouse Gas Emissions

Data Type: XLSX
Source: [3]
This is focused on emissions from transportation, this dataset details trends and breakdowns by sector, highlighting the carbon cost of logistics in the supply chain.

2.4 U.S. Greenhouse Gas Emissions from Domestic Freight Transportation

Data Type: CSV
Source: [4]
This dataset specializes in freight transportation emissions, offering detailed temporal data on the carbon impact of freight activities within domestic U.S. logistics.

2.5 Data Licensing and Compliance

All datasets are released under open-data licenses CC0 and public domain, permitting use for analysis and re-

porting. Compliance is ensured through proper citation and documentation practices.

2.6 Data Structure and Quality

The datasets come in structured formats (CSV, XLSX), with well-defined variables. While the quality is generally high, preprocessing addressed missing values and inconsistencies to ensure reliable analysis.

3 Data Pipeline

In this study, I employed an automated ETL (Extraction, Transformation, and Loading) data pipeline to process and analyze greenhouse gas (GHG) emissions data across various supply chain sectors in the United States. The pipeline automates the extraction of datasets, cleans and transforms the data, and loads it into a structured format for analysis. The process ensures that the data is readily available for generating insights into the variations in GHG emissions across sectors.

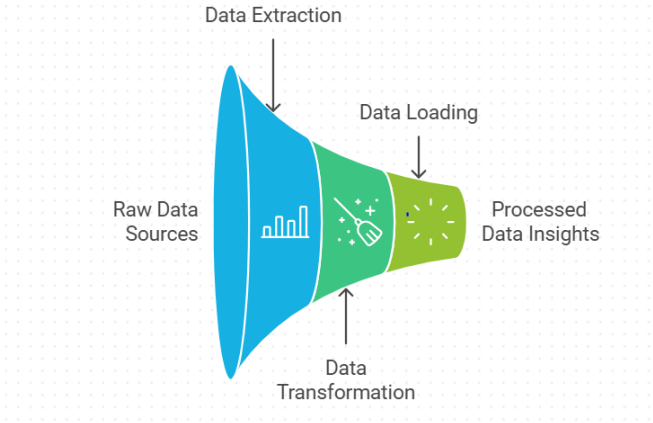

Figure 1: ETL data pipeline architecture

3.1 Data Extraction

The extraction phase starts by downloading the datasets from reliable online sources. Using the `requests` library, the data files—stored in CSV and Excel formats—are pulled from the web. These datasets are then stored locally in the `data` directory. Once extracted, the data is loaded into pandas `DataFrame`s, making it ready for further processing and analysis.

3.2 Data Transformation and Cleaning

Following the extraction, the next phase involves transforming and cleaning the data to ensure its quality and usability:

- Removing Unnecessary Columns: Irrelevant columns, such as indices or non-essential metadata, were removed to focus on the core data.
- Handling Missing Data: Missing values were addressed by applying a combination of deletion and imputation.
- Reformatting: I ensured that all decimal-related columns were adjusted to the required precision levels, facilitating accurate computations and analyses.
- Standardization: To standardize the data, columns containing emission values were adjusted at certain level, and values were converted into consistent units.

For these transformations, I relied heavily on pandas, which offered powerful tools for data cleaning, manipulation, and restructuring.

3.3 Loading Data into the Sink

Following the data cleaning and transformation processes, the processed data is systematically loaded into two primary storage formats for subsequent analysis:

- **SQLite Database (emission.db):** The intermediate data, representing the cleaned but unprocessed data, is stored in the SQLite database `emission.db`. This database acts as a temporary staging area, where data is organized for initial stages of analysis.
- **SQLite Database (insights.db):** The final, transformed data is loaded into the `insights.db` database, which serves as the data sink. This database stores the processed data in a structured, organized manner, with clearly defined schema and standardized naming conventions to support detailed analysis and easy access.

This structured approach ensures that the data is well-organized, easily accessible, and ready for analysis. During the development of this pipeline, one **challenge** was managing the temporary storage of downloaded files. To address this, I implemented a flag-based approach using the `--use-cache` option. This approach ensures that files are only re-downloaded if necessary, optimizing storage space by reusing previously downloaded files when available.

However, A major **limitation** of this current data pipeline is the fact that it is not dynamic with changing data. It does not support updates or changes in the structure of the data, including emission factors or NAICS codes. It also lacks control over versions of datasets, overwriting previous versions without tracking any updates.

3.4 Data Analysis

The processed data, now stored in an SQLite database `insights.db` , is ready for in-depth analysis. With the structured dataset, I can query specific requirements through available data, such as:

- Emissions factors for various industries (e.g., NAICS codes)
- Historical GHG emissions trends (1990-2022)
- Sectoral emissions breakdown by commodity or transportation-related activities

By leveraging SQL queries and pandas, I can extract meaningful insights and analyze how GHG emissions vary across different supply chain sectors in the U.S., providing valuable information to drive sustainability efforts in the industry.

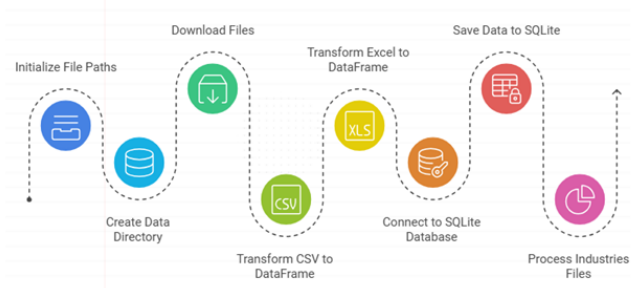

Figure 2: Pipeline - U.S. Supply Chain Carbon Footprint

4 Results

4.1 Output Data of the Pipeline

The pipeline outputs a structured SQLite database containing processed greenhouse gas (GHG) emissions data across U.S. supply chain sectors. Key tables include in Figure 3.

4.2 Why SQLite?

SQLite was chosen for its efficient architecture, seamless Python integration, and portability, enabling high-performance management of medium-sized datasets without complex infrastructure or configurations.

Figure 3: Tables in insight.db

4.3 Enhancements

Future advancements could augment the pipeline's functionality:

- **Adaptive Normalization**: Implementing adaptive normalization algorithms would ensure dynamic data scaling as new inputs are ingested.
- **Real-time Data Ingestion**: Integrating real-time data streams would enable continuous, up-to-the-minute analysis of GHG emissions.
- **Expanded Variable Integration**: Incorporating additional parameters, such as energy consumption metrics and carbon sequestration technologies, would enhance the granularity of the environmental impact model.
- **Geospatial Analytics**: Embedding geospatial data analytics would allow for spatially targeted emissions assessment, identifying regional hotspots for focused mitigation interventions.

5 Conclusion

The data pipeline efficiently handled the ETL operation of GHG emissions data from sources using Python and ETL architecture, it processed large datasets, providing insights into sector-specific emissions within the U.S. supply chain. Overall, the pipeline serves as a solid foundation for understanding and mitigating the environmental impact of U.S. supply chains.

References

1. Supply Chain GHG Emission Factors v1.3, *Data.gov*. `https://data.gov`
2. Transportation-Related GHG Emissions, *Bureau of Transportation Statistics*. `https://bts.gov`
3. Greenhouse Gas Emissions in the U.S., *Net0*. `https://net0.com/blog`
4. Carbon Footprints in Supply Chains, *FAO*. `https://openknowledge.fao.org`
5. M. Alam, "Predictive Analytics for Sustainable Supply Chains," *ResearchGate* `https://researchgate.net`
6. Efficient Strategies for Supply Chain Carbon Reduction, *arXiv*. `https://arxiv.org/abs/2404.16863`
7. U.S. Supply Chain Carbon Footprint Analysis, *ScienceDirect*. `https://sciencedirect.com`