# ****North American Freight Data Analysis****

**Find the link to my project with well-detailed README below**

https://github.com/JohnEvansOkyere/Trans\_BorderFreight\_Data\_Analysis

**Prepared by:** John Evans Okyere  
**Methodology:** CRISP-DM (Cross-Industry Standard Process for Data Mining)  
**Tools Used:** Python (Pandas, Matplotlib, Seaborn), Jupyter Notebook

## ****1. Business Understanding****

The goal of this project is to provide actionable insights into freight transportation trends across North America, specifically between the U.S., Canada, and Mexico. This includes identifying the most utilized transport modes, ports, routes, and trade patterns using data from the Bureau of Transportation Statistics (2020–2024).

### ****Key Business Objectives****

1. **Identify Seasonal Freight Patterns**  
   Determine which seasons and months record the highest shipment volumes and revenues to optimize resource allocation and pricing strategies.
2. **Understand the Relationship Between Shipment Weight and Trade Value**  
   Analyze how shipment size correlates with economic value to inform pricing, insurance, and risk management.
3. **Uncover Dominant Freight Movement Patterns by Mode, Region, and Time**  
   Identify the most frequently used transportation modes and how their usage varies across states and over time to guide infrastructure and fleet investments.
4. **Quantify Revenue Contribution by Transportation Mode**  
   Determine which modes generate the most revenue to focus commercial and operational efforts on the most profitable segments.
5. **Assess Operational Efficiency and Cost Inefficiencies**  
   Pinpoint transportation modes and routes with the highest cost per weight and freight charges to drive cost-reduction initiatives.
6. **Prioritize Commodities for Supply Chain and Policy Planning**  
   Identify top commodities by shipment count and trade value to support supply chain prioritization and risk management.
7. **Analyze Cross-Border Trade Patterns and Modal Utilization**  
   Understand key trade flows with Canada and Mexico and the dominant modes used, enabling tailored cross-border strategies.
8. **Evaluate Seasonal Trends in Freight Movement and Costs by Mode**  
   Reveal how freight volumes and costs fluctuate seasonally across modes to inform budgeting, contracting, and resource planning.
9. **Detect Economic Disruption Impacts**  
   Monitor the correlation between trade value and shipment weight over time to identify and respond to economic shocks.
10. **Optimize Infrastructure Utilization and Address Bottlenecks**  
    Identify top ports and districts by volume and revenue, and detect potential bottlenecks for targeted infrastructure investment.
11. **Improve Containerization Efficiency**  
    Assess the impact of containerization on cost and trade value across modes and routes to optimize equipment and operational strategies.

## ****2. Data Understanding****

The data was sourced from the Bureau of Transportation Statistics and includes:

* Monthly transborder freight data (2020–2024)
* Modes of transport (e.g., Truck, Vessel, Rail, Air, Pipeline)
* Geographic information (Origin/Destination State, Port District)
* Trade values, weight (kg), containerization, and trade type

**Files Used:** dot1, dot2, dot3  
**Volume:** Millions of freight records across multiple years

## ****3. Data Preparation****

Data was preprocessed using custom scripts for:

* **Standardization:** Harmonized column names and file formats
* **Merging:** Combined monthly files into unified datasets by file type
* **Mapping:** Converted codes into descriptive labels (states, ports, modes, etc.)
* **Cleaning:** Removed duplicates, handled missing values, and filtered erroneous entries
* **Period Creation:** Combined Year and Month into a usable Period column for time series

## ****4. Exploratory Data Analysis (EDA)****

### Questions Answered:

| # | Analytical Question |
| --- | --- |
| 1 | What transportation mode carries the most freight over time? |
| 2 | Which seasons or months recorded the highest weight and trade value? |
| 3 | Which ports are most utilized by shipment volume? |
| 4 | What is the relationship between shipment weight and trade value? |
| 5 | How does containerization impact average trade value? |

## ****5. Key Insights****

### ****1. Seasonal Volume and Revenue Performance****

**Freight Volume Trends:**

* Shipment volumes show clear seasonal fluctuations, with distinct peaks and troughs across months and years.
* The highest shipment volumes (by weight) typically occur during specific months, indicating strong seasonality in freight demand.

**Revenue Trends:**

* Trade value (revenue) also exhibits seasonality, often peaking in the same periods as shipment weight, but with some divergence—suggesting that high volume does not always equate to high revenue.

**Business Implication:**  
Seasonal planning is essential. Aligning resources and pricing strategies with peak periods can improve capacity utilization and profitability.

### ****2. Relationship Between Shipment Weight and Trade Value****

**Correlation Analysis:**

* Scatter plots reveal a positive but non-linear relationship between shipment weight and trade value.
* Some shipments with low weight have high trade value (likely high-value goods), while others with high weight have moderate value (bulk commodities).

**Business Implication:**  
Pricing, insurance, and security strategies should consider both weight and value, not just one metric.

### ****3. Freight Movement Patterns****

**Dominant Modes:**

* Certain transportation modes (e.g., Truck, Rail) dominate the shipment landscape.
* The top 10 U.S. states by freight volume are visualized, with mode distribution varying by state.

**Temporal Trends:**

* The use of different modes changes over time, with some modes gaining or losing share in response to market or policy shifts.

**Business Implication:**  
Investments in infrastructure and fleet should be mode- and region-specific, reflecting actual usage patterns.

### ****4. Freight Movement Revenue by Mode****

**Revenue Contribution:**

* A small number of modes contribute the majority of freight revenue.
* Modes with less than 2% revenue share are grouped as "Other Modes" for clarity.

**Business Implication:**  
Focus commercial and operational efforts on the most lucrative modes.

### ****5. Operational Efficiency****

**Cost per Weight by Mode:**

* Significant variation exists in average cost per weight across modes.
* Some modes are 3–5x more expensive per unit weight.

**Most Inefficient Mode-Route Combinations:**

* The top 3 most inefficient mode-route pairs are identified, highlighting where costs are highest.

**Business Implication:**  
Route and mode optimization can yield substantial cost savings.

### ****6. Commodity Analysis****

**Top Commodities:**

* The most frequently shipped commodities (by count) and those with the highest total trade value are identified.
* There is not always overlap between high-volume and high-value commodities.

**Business Implication:**  
Supply chain prioritization and risk management should focus on both high-volume and high-value commodities.

### ****7. Cross-Border Trade Analysis****

**Country-Mode Patterns:**

* The U.S. trades with Canada and Mexico using different dominant modes.
* Stacked bar charts show mode distribution by country.

**Most Used Modes:**

* A table lists the most used modes for cross-border freight by total weight.

**Business Implication:**  
Cross-border strategies should be tailored to each trading partner’s modal mix.

### ****8. Seasonal Trends by Mode****

**Freight Movement:**

* Seasonal trends in shipment volumes are visualized for each mode, revealing which modes are most sensitive to seasonality.

**Freight Costs:**

* Seasonal cost trends by mode show when and where freight charges spike.

**Business Implication:**  
Seasonal contracts and resource allocation should be mode-specific.

### ****9. Economic Disruption Impact****

**Normalized Trends:**

* Freight weight and trade value are normalized and plotted together, showing how their trends align or diverge over time.

**Business Implication:**  
Monitoring these trends can help detect economic shocks or disruptions early.

### ****10. Infrastructure Utilization****

**Top Ports by Volume and Revenue:**

* The top 10 ports/districts are identified by both shipment weight and trade value.

**Potential Bottlenecks:**

* Ports with the highest average shipment weight are flagged as potential bottlenecks.

**Business Implication:**  
Targeted investment in high-volume or bottleneck ports can improve network efficiency.

### ****11. Containerization Efficiency****

**Underutilized but Cost-Effective Routes:**

* Some routes are underutilized yet have low cost per weight, representing opportunities for growth.

**Containerization Impact:**

* The average trade value by container type and mode is visualized, showing where containerization adds the most value.

**Most Cost-Efficient Modes per Country:**

* For each country, the most cost-efficient mode is identified.

**Business Implication:**  
Promote containerization and shift volume to cost-effective routes and modes.

## ****6. Actionable Recommendations****

* **Optimize Transportation Mode Selection:**  
  Automate mode selection for cost efficiency (15–25% cost reduction).
* **Expand High-Performing Port Capacity:**  
  Invest in top ports and develop secondary facilities (20–30% throughput improvement).
* **Implement Seasonal Planning:**  
  Pre-position equipment and negotiate seasonal pricing (10–15% better utilization).
* **Optimize Container Usage:**  
  Prioritize containerization where most beneficial (12–18% efficiency improvement).
* **Streamline Cross-Border Operations:**  
  Develop corridor-specific strategies for faster, more efficient trade flows.

## ****7. Deployment and Reproducibility****

To reproduce the results:

# Step 1: Aggregate raw files

python preprocess/combining.py

# Step 2: Apply mapping

python preprocess/renaming\_mappin.py

# Step 3: Launch Jupyter and run analysis

jupyter lab

All notebooks, preprocessing scripts, and cleaned datasets are stored and versioned in the Git repository.

## ****8. Future Enhancements****

* **Real-Time Data Integration** via APIs
* **ML Forecasting** for freight demand and seasonal trends
* **GIS Mapping** for spatial visualization
* **Carbon Emission Calculators** by route and mode