SCM – supply chain management

Forecast reconciliation – coherent and accurate

Repository, slide deck, 5-minute presentation

**Data cleaning summary:**

✅ **Script Created**: clean\_supply\_chain\_data.py - A professional, human-friendly data cleaning tool✅ **Data Processed**:

* 32,065 rows processed with 100% quality score
* Added 4 quality monitoring flags for risk analysis
* Corrected 96,195 binary field values
* Validated GPS coordinates (Southern California region)
* Validated temporal data (Jan 2021 - Aug 2024, 1,336 days)
* Sorted data chronologically

✅ **Output Files**:

* cleaned\_supply\_chain\_data.csv - Your cleaned dataset ready for analysis
* data\_cleaning\_report.txt - Simple, professional quality report

**What the Script Does**

The cleaning pipeline:

1. Loads and validates data types (timestamps, numeric, binary fields)
2. Checks for and handles missing values intelligently
3. Removes duplicate records
4. Validates and corrects value ranges (GPS coordinates, scores 0-1, levels 0-10, etc.)
5. Validates geospatial data for Southern California
6. Sorts data chronologically by timestamp
7. Detects outliers (without removing them - they may be valid extreme events)
8. Adds quality monitoring flags for high-risk situations
9. Filter out date from Jan 2021 – Jan 2024

**Cleaning Actions Performed:**

Bottlenecks analysis summary:

Interpretation:

- 'Affected %' = % of total shipments affected by this bottleneck

- 'Late %' = % of affected shipments that are late (ETA variation > 0)

- Compare 'Late %' to overall late rate of 100.0% to see impact

Threshold Methodology:

- Most bottlenecks use 75th percentile (worst 25% of cases)

- This ensures we capture meaningful outliers, not just average performance

- Binary variables (equipment, cargo) use actual status (0 = problem)

**1. Late Delivery (bn\_late)**

* **Threshold**: ETA variation > 3 hours
* **Meaning**: Deliveries that are severely late (more than 3 hours behind schedule)
* **From terminal output**: This is the most common bottleneck

**2. No Equipment (bn\_no\_equip)**

* **Threshold**: Handling equipment availability = 0
* **Meaning**: No forklifts, loading equipment, or handling tools available
* **Impact**: Critical operational constraint affecting 72.3% of deliveries (from earlier analysis)

**3. High Traffic (bn\_traffic)**

* **Threshold**: Traffic congestion level > 7 (out of 10)
* **Meaning**: Severe traffic conditions on the route
* **Impact**: Significantly delays deliveries

**4. Slow Loading (bn\_slow\_load)**

* **Threshold**: Loading/unloading time > 75th percentile (top 25% slowest)
* **Meaning**: Loading operations taking longer than normal
* **Impact**: Delays compound with other bottlenecks

For data analysis:

* Unweighted correlation: computes Pearson r over all rows equally. Here that’s r between per-day stress\_index and binary is\_late. Each day contributes the same, regardless of how many shipments happened that day.
* Weighted correlation: first aggregates by stress\_index level to get a late rate per level, then correlates stress level vs late rate, weighting each level by its row count (in your current code). This:
* Uses smoothed group rates (reduces variance vs row-level binary outcomes).
* Gives more influence to stress levels that occur more often.
* Ignores within-level dispersion.