Data Cleaning & Validation Framework

This document outlines the structured approach used to clean, validate, and prepare datasets for analysis within the Inventory Operations Enhancement initiative. The framework is designed to reflect real-world analyst practice, balancing executed transformations with validation logic and scalability considerations.

What Was Cleaned, Validated, and Designed For

This data preparation work followed a disciplined separation between hands-on cleaning, logical validation, and framework design for real-world variability. This distinction ensures transparency around what was directly executed in this project versus how the approach would scale in a production environment.

What I Cleaned (Directly Executed Transformations)

The following steps were actively performed during dataset preparation to ensure analytical readiness:

- Standardized date formats across order, inventory, and supplier datasets to enable time-based analysis.
- Normalized categorical fields such as SKU IDs, warehouse codes, supplier names, and product categories to ensure consistent joins.
- Corrected data type mismatches across numeric, date, and text fields to prevent calculation and aggregation errors.
- Removed exact duplicate records caused by repeated synthetic generation while preserving legitimate multi-record transactions.
- Flagged null or blank values in critical operational fields instead of auto-imputing values without context.

These actions focused on structural cleanliness and consistency rather than altering business meaning.

What I Validated (Analyst-Side Logical and Business Rule Checks)

Validation focused on confirming that the cleaned data behaved consistently with expected operational logic. This validation was performed using business rules and internal consistency checks, not stakeholder sign-off.

Examples include:

- Verified revenue consistency using unit price and quantity.
- Confirmed inventory availability logic (ATP) aligned with commitments.
- Ensured delivery dates occurred after order dates.
- Validated supplier lead times were mathematically consistent with order and receipt dates.
- Checked referential integrity across all datasets.

Any records that violated these rules were flagged and reviewed, not automatically corrected. In a real-world engagement, unresolved ambiguity identified during this validation phase would be escalated to business stakeholders before final decisions were made.

What the Framework Was Designed to Handle (If Data Were Messier)

Although the datasets exhibited moderate-to-high inconsistencies, the overall approach was intentionally designed to scale to more complex real-world scenarios without changing methodology.

The framework anticipates:

- Late-arriving or missing operational records.
- Conflicting values across multiple system extracts.
- Partial inventory visibility due to execution timing gaps.
- Duplicate or overlapping records from parallel processes.
- Incomplete supplier performance data.

This ensures the approach remains robust even when applied to production-grade operational data.

Business Rules Used For Validation

Inventory and Availability Rules

- Available-to-promise must not exceed total on-hand inventory.
- Negative available inventory indicates execution timing gaps, not normal state.
- Inventory quantities must be non-negative at warehouse level snapshots.

Order and Revenue Rules

- Revenue = unit price × order quantity.
- Order quantity must be greater than zero.
- Order date must exist for any revenue-generating transaction.
- Delivery date must be greater than or equal to order date.

Supplier and Inbound Rules

- Lead time = delivery date − order date.
- Lead time must be non-negative.
- On-time delivery status must align with promised vs actual delivery dates.
- Supplier IDs must be consistent across inbound and inventory tables.

Cross-Dataset Integrity Rules

- Warehouse codes must be consistent across inventory and fulfillment tables.
- Category and product hierarchy must remain stable across reporting periods.

This list is realistic, defensible, and exactly what analysts are expected to articulate.