

Key Findings in Data

Domain Table	Field	Requirement / Constraint	Cleansing Rules Applied / Rationale
Customer Dim	customer_id	Must be unique, numeric, non-null	Enforce PK constraint, remove duplicates
	first_name	Alphabetic only (^[A-Za-z .'-]+\$), min length 2	Clean special chars (N!cole -> Nicole), remove digits
	last_name	Alphabetic only, min length 2	Remove digits (R0berts -> Roberts)
	age	Integer 18–100	Drop outliers, non-numeric values
	country	Valid ISO country name	Standardize (e.g., U.S.A. -> USA), remove invalid chars
Order Fact	order_id	Must be unique, numeric, non-null	Enforce PK constraint
	amount	DECIMAL(15,2), must be > 0	Remove negatives, enforce numeric check
	item	Alphanumeric + limited punctuation (()-'.)	Remove invalid chars (Shoe#123 Shoe123), standardize casing
	customer_id	FK customer_dim.customer_id	Block orphan records
Shipping Dim	shipping_id	Must be unique, non-null	Enforce PK constraint
	status	Enum: {Pending, Shipped, Delivered, Cancelled}	Reject/clean invalid values
	customer_id	FK customer_dim.customer_id	Enforce referential integrity
All Tables	Timestamps / Metadata	Capture load_date, source_system, quality_flag	Enable lineage and auditing

Areas With No Issues

- Missing Value Analysis - No missing IDs, names, ages, countries, amounts, or statuses.
- Referential Integrity Checks - No orphaned Orders or Shipping records.
- Data Consistency - No duplicate IDs found.
- Business Rule Validation - No negative order amounts, no invalid ages (<18 or >100).

Cleansing Rules Derived

1. Customer Names:

- Must be alphabetic only.
- Length ≥ 2 .
- Remove digits/special chars (R0bert → Robert, N!cole → Nicole).

2. Order Items:

- Remove special characters (#, @, \$).
- Keep alphanumeric + basic punctuation (() ' - .).
- Standardize casing (Title Case).

3. Country Names:

- Map to ISO standard list (e.g., U . S . A . -> USA).
- Remove invalid chars.

4. Ages:

- Keep only integers in [18–100].
- Outliers below/above flagged for manual review.

5. Order Amounts:

- Must be numeric > 0 .
- Non-numeric values discarded or corrected.