Assumptions for Supplier Data Cleaning

Schema & Joining

- Canonical schema is fixed to: source, article_id, material, grade, quality_choice, finish, thickness_mm, width_mm, weight_kg, quantity, rp02, rm, ag, ai, reserved, description.
- Join strategy: vertical concatenation (stack/append) into one table; no cross-vendor key existed to perform a relational join.
- Traceability: a source column is added (supplier1 / supplier2) to preserve provenance of each row.
- No de-duplication was performed because no reliable cross-file unique key existed; if needed, a downstream rule such as (material, thickness_mm, width_mm, weight_kg) can be applied.

Units & Parsing

- Thickness/width are millimeters (*_mm).
- Weight is kilograms (weight_kg).
- Mechanical properties (rp02, rm, ag, ai) are left as provided (no unit conversions assumed).
- Header normalization: column names converted to snake_case, and unit markers in parentheses incorporated (e.g., Thickness (mm) → thickness_mm).
- Numeric coercion: robust parsing accepts messy strings and European formats (decimal comma, thousand separators). First numeric token is taken if multiple appear.
- Quantity dtype: cast to nullable integer (Int64) only if all non-null values are whole numbers; otherwise kept as numeric floats.

Field-Level Rules

- Weight: Supplier 1 Gross weight (kg) treated as weight_kg. If both net/gross existed, gross took precedence.
- Thickness/Width: used explicit numeric columns when present. For Supplier 2, parsed t x w patterns (e.g., 1,50x1250) from material if missing.
- Finish: standardized to pickled, unpickled, oiled, galvanized, bright, brushed. Priority cues from explicit finish; else inferred from description/material.
- Quality/Choice (Supplier 1): mapped variants to ordinals (A/1/i/prime→1st; B/2/ii→2nd; C/3/iii→3rd; D/4/iv→4th). Ambiguous tokens left as-is.
- Grade: Supplier 2: extracted steel grade tokens from material (DX51D, S235JR, C100S, DC01, DD11, etc.). Backfilled from description if still missing. Material text itself not altered.
- Reserved: normalized textual flags (yes/ja/1/true → True; no/nein/0/false → False). If missing and description contains 'reserv…', inferred True; otherwise left NaN.
- Description: whitespace normalized; literal 'nan' strings removed; content otherwise preserved.

Missing/Invalid Data Handling

- No row deletion solely due to missing thickness_mm/width_mm/grade; such rows are retained.
- No outlier filtering and no imputations beyond parsing/inference steps.
- Non-numeric or unparsable numeric fields are set to NaN.

Language & Text

- German terms are supported for finish and quality mappings (e.g., 'gebeizt/ungebeizt/geölt', 'erst/zweit/dritt').
- Coating indicators like '+Z' imply galvanized finish.

Limits / Non-assumptions

- No business logic applied to convert gross ↔ net or adjust for packaging.
- No attempt to reconstruct article IDs for Supplier 1.
- No inference of width/thickness from textual patterns beyond implemented regex.