

Data Cleaning Report

messy_vendors.csv · Generated 7/28/2026, 6:24:53 PM

Summary

15 rows in → 7 rows out

Vendor contract list (15 rows) with inconsistent vendor name spellings, mixed date formats (including an Excel serial), unformatted/dot/dash phone numbers, currency-symbol/comma-mixed amounts, two fully empty rows, and one exact duplicate row.

Applied changes

- Whitespace & ghost-char cleanup (cells) — 6
- Date format unified — 10 (Contract Date → YYYY-MM-DD)
- Phone format unified — 8 (Phone)
- Numbers & amounts normalized — 13 (Amount)
- Name variants unified — 6 (Vendor)
- Empty rows removed — 2
- Duplicate rows removed — 6

Review notes

1) Canonical vendor names were chosen as the shorter/first-seen variant (e.g. 'Acme Corp'); confirm whether legal names ('Acme Corporation') are preferred. 2) 'Contract Date' value 45291 is an Excel serial that converts to 2023-12-31, which conflicts with the other Initech row (2024-03-15) — needs human verification. 3) After vendor/date/phone/amount normalization, several near-duplicate rows (e.g. the three Acme rows, two Globex rows) will become identical and be collapsed by dedupeRows — confirm these are truly the same contracts, not separate ones. 4) trim removes leading/trailing spaces but not internal double spaces ('John Kim'); that one value may still differ from 'John Kim' unless whitespace is also collapsed.

Every change is traceable via the list above · leetei studio · leetei.com