

KitForma.

PRACTICAL FIELD GUIDE

CSV Cleaning Without Losing Meaning

An original workbook for exact identifiers, dates, currencies, duplicates, and reliable exports

01

02

03

SAMPLE EDITION

ENGLISH / 1.0 / 2026

How to use this guide

People preparing small CSV or values-only spreadsheet exports for another person or system.

Book by KitForma. Ten practical chapters follow one original eight-record dataset through parsing, explicit rules, quarantine, deduplication, export, and an independently checkable capstone. Includes source fixtures, expected results, and an answer key beyond the freely available public tool guides.

- Parse quoted commas, quotation marks and multiline fields correctly.
- Preserve 0012 and 9007199254740993 as exact strings.
- Resolve declared date and decimal conventions without guessing unknown values.
- Reconcile accepted, quarantined and duplicate rows.
- Produce and reopen an export suited to its actual consumer.

Scope and limitations

- Not a financial audit, currency conversion service, or automatic source correction.
- No universal spreadsheet formula-safety guarantee; intended consumers require their own reopen checks.
- Fixtures are fictional and small; production data may require additional field and volume rules.
- The exercises cover small datasets; test larger files and your receiving software before using them in production.

Published by KitForma. Edition 1.0. Sources checked 1 October 2026. People and businesses in worked examples are fictional. This is an independent publication, not endorsed by the platforms mentioned.

Copyright 2026 KitForma. Usage, support and refund terms are included in LICENSE.txt in the purchased package.

Contents

How to use this guide	2
Contents	3
Write a cleaning contract before touching values	4
Sources and tools	6

01

Write a cleaning contract before touching values

Data cleaning is a series of decisions about meaning. Removing every unusual character or converting every number-shaped cell can make a table look tidy while damaging it. Begin with the destination: what will the next person or program do with this file? For this workbook, a fictional workshop shop needs a checked order extract, a separate unresolved-row list, and a small review report. It does not need invented dates, guessed currency conversions, or a spreadsheet that silently changes identifiers.

Keep the original export unchanged. Work on a named copy, retain the source row reference, and record each rule before applying it. The supplied `orders-raw.csv` contains eight fictional records and nine columns. It intentionally includes a quoted comma, embedded quotation marks, a multiline note, a duplicate, an impossible date, an ambiguous date, a missing amount, and a harmless formula-shaped note. These are controlled examples, not real customer information.

Define the unit of a record. In our exercise, each record describes one order total, not one item price. Quantity gives the count associated with that order, but the amount is already the line total. Multiplying amount by quantity would therefore double count. Define the business key as `order_id`. Define a repeat as an otherwise identical record with the same order identifier; exclude `row_ref` from that comparison because it describes provenance rather than the order.

The contract permits zero quantity and zero amount for an explicitly marked sample. It requires a known date interpretation, a real calendar date, a present amount, and one of USD, EUR, or TRY. It preserves order identifiers exactly as text. It does not sum across currencies. Unresolved

records go to quarantine with reasons, and duplicates go to a separate ledger. Neither category disappears from the accounting of input rows.

Book by KitForma. The book's original contribution is this connected fictional dataset, its rule decisions, workbook exercises, expected outputs, and capstone. Public tool pages and the primary references remain available independently. The exercises can be carried out with a suitable CSV reader, a spreadsheet imported with explicit text types, or a script using a real CSV parser. No purchase or particular programming language is required.

State the contract

Write five rules for the supplied file: identifier type, date interpretation, amount meaning, duplicate definition, and treatment of unresolved rows. Include one sentence explaining why a blank amount cannot be replaced with zero merely to complete the table.

Sources and tools

[S1] RFC 4180: Common Format and MIME Type for CSV Files

<https://www.rfc-editor.org/rfc/rfc4180>

Checked: 2026-10-01

[S2] Microsoft: Keeping leading zeros and large numbers

<https://support.microsoft.com/en-us/excel/keeping-leading-zeros-and-large-numbers>

Checked: 2026-10-01

[S3] RFC 8259: The JSON Data Interchange Format, section 6

<https://www.rfc-editor.org/rfc/rfc8259>

Checked: 2026-10-01

[S4] OWASP: CSV Injection

https://community.owasp.org/attacks/CSV_Injection

Checked: 2026-10-01

CSV & Excel column mapper

Optional local column mapping, row validation and text-cell XLSX or report export; manual business decisions remain necessary.

<https://kitforma.com/en/csv-xlsx-mapper/>

CSV & Excel chart report

Optional chart only after cleaning, with currencies kept separate; not a validation or currency-conversion substitute.

<https://kitforma.com/en/csv-chart-report/>

Inside the complete guide

Parse records rather than splitting visible lines

Protect identifiers before assigning types

Resolve dates through declared meaning

Handle amounts, currencies, blanks, and zeros

Clean text deliberately and deduplicate by a key

Choose an export for its actual consumer

Reconcile every row and test the round trip

Capstone: deliver the fictional workshop extract

Answer key and diagnostic repairs