Portfolio sample: synthetic demo data, not client work.
This sample was produced with AI tools and checked by an automated QA step against an internal demo brief.

Data cleanup sample: validation report and change log

300 raw rows in, 272 clean rows out: 8 unrecoverable rows dropped with reasons, 8 exact and 12 near duplicates removed, 30 validation checks with 0 failed.

Validation report - PASSED

Generated 2026-10-18 09:31 UTC by the validation step (pandera 0.34.1). Demo data is synthetic.

- Rows validated: **272**
- Columns: 14 (as shipped in the clean CSV, including the data_note column; the internal _fixes helper column is excluded)
- Checks: **15 rule checks plus 2 uniqueness and 13 type/nullability checks** (30 total, each applied to every row)
- Checks failed: **0**; failure cases: **0**

Rule checks

- PASS - _raw_row: greater_than_or_equal_to(1)
- PASS - Order ID: str_matches('^ORD-\d{5}$')
- PASS - Order Date: greater_than_or_equal_to(2020-01-01 00:00:00)
- PASS - Order Date: less_than_or_equal_to(2030-12-31 00:00:00)
- PASS - Customer: str_length(2, None)
- PASS - Customer Phone: str_matches('^\+\d{1,3}-\d{3}-\d{3}-\d{4}$')
- PASS - Region: isin(['Central', 'East', 'North', 'South', 'West'])
- PASS - Category: isin(['Accessories', 'Electronics', 'Furniture', 'Office Supplies'])
- PASS - Units: in_range(1, 10000)
- PASS - Unit Price: in_range(0.01, 100000)
- PASS - Revenue: greater_than_or_equal_to(0)
- PASS - derived_from_revenue: isin(['Units', 'Unit Price'])
- PASS - table: revenue_equals_units_x_price
- PASS - table: one_category_per_product
- PASS - table: no_duplicate_records

Uniqueness checks

- PASS - _raw_row: values unique
- PASS - Order ID: values unique

Type / nullability checks

- PASS - _raw_row: dtype int64, not null
- PASS - Order ID: dtype str, not null
- PASS - Order Date: dtype datetime64[ns], not null
- PASS - Customer: dtype str, not null
- PASS - Customer Phone: dtype str
- PASS - Region: dtype str, not null
- PASS - Product: dtype str, not null
- PASS - Category: dtype str, not null
- PASS - Units: dtype int64, not null
- PASS - Unit Price: dtype float64, not null
- PASS - Revenue: dtype float64, not null
- PASS - Sales Rep: dtype str, not null
- PASS - derived_from_revenue: dtype str

CHANGELOG - data cleanup run

Auto-generated on 2026-10-18 09:31 UTC from the pipeline run log.

- Input: `raw_sales_messy_SYNTHETIC.csv` (md5 `d22dd6234790d6cb6932aaf7348343f3`)
- Tool versions: polars 2.0.0, duckdb 1.5.6, splink 5.0.0, pandera 0.34.1, ydata-profiling 4.18.4, openpyxl 3.1.5, pandas 2.3.3, Python 3.13.5

Step 1: profile raw data (2026-10-18 09:31:29 UTC)

Profiled raw input: 300 rows x 11 columns -> profile_before.html

- rows: 300
- columns: Order ID, Order Date, Customer, Customer Phone, Region, Product, Category, Units, Unit Price, Revenue, Sales Rep
- blank_cells_by_column:
  - Customer: 4
  - Customer Phone: 28
  - Region: 8
  - Category: 5
  - Units: 2
  - Unit Price: 3
  - Revenue: 31
  - Sales Rep: 3
- exact_duplicate_rows: 8

Step 2: clean (2026-10-18 09:31:30 UTC)

Standardised 300 rows; kept 292 (6 with Units/Price derived from Revenue), dropped 8 unrecoverable rows

- rows_in: 300
- rows_out: 292
- rows_dropped: 8
- rows_derived_from_revenue: 6
- drop_reasons:
  - Order Date: impossible date: 6
  - Order Date: unparseable date: 1
  - Unit Price: missing required value, not derivable from Revenue: 1
- date_styles_seen:
  - YYYY-MM-DD (ISO): 127
  - DD/MM/YYYY: 71
  - Month D YYYY (March 4 2026): 33
  - DD-Mon-YYYY (04-Mar-2026): 25
  - Mon D, YYYY (Mar 4, 2026): 15
  - YYYY-MM-DD hh:mm:ss (ISO date-time; time dropped): 13
  - DD.MM.YYYY: 9
- counts_all_raw_rows:
  - Category_values_mapped: 131
  - Customer_text_standardised: 92
  - Product_text_standardised: 10
  - Region_values_mapped: 135
  - Sales Rep_text_standardised: 88
  - Unit Price_derived_from_revenue: 3
  - Unit Price_text_to_number: 133
  - Units_derived_from_revenue: 3
  - Units_text_to_number: 112
  - date_fixed: 166
  - date_impossible: 6
  - date_ok_iso: 127
  - date_unparseable: 1
  - ids_normalised: 25

    - phone_blank: 28
    - phone_ok: 272
    - phones_reformatted: 272
- counts_kept_rows:
    - Category_inferred_from_Product: 5
    - Customer_filled_default: 4
    - Region_inferred_from_Sales Rep: 8
    - Sales Rep_filled_default: 3
    - raw_revenue_blank: 30
    - raw_revenue_wrong: 32

Step 3: deduplicate (2026-10-18 09:31:31 UTC)

Removed 8 exact + 12 near duplicates (Splink, p>=0.95); 272 rows remain

- rows_in: 292
- exact_duplicates_removed: 8
- near_duplicates_removed: 12
- rows_out: 272
- candidate_pairs_scored: 14
- borderline_pairs_for_review: 0
- order_id_conflicts_remaining: 0

Step 4: validate (2026-10-18 09:31:31 UTC)

pandera schema PASSED: 15 rule checks plus 2 uniqueness and 13 type/nullability checks on 272 rows; 0 checks failed, 0 failure cases

- status: passed
- rows: 272
- checks_summary: 15 rule checks plus 2 uniqueness and 13 type/nullability checks
- checks_total: 30
- rule_checks: 15
- uniqueness_checks: 2
- type_checks: 13
- checks_failed: 0
- failure_cases: 0
- clean_csv: sales_clean_SYNTHETIC.csv

Step 5: profile clean data (2026-10-18 09:31:35 UTC)

Profiled cleaned output: 272 rows x 13 columns -> profile_after.html

- rows: 272
- blank_cells: 288
- blank_cells_by_column:
    - Customer Phone: 22
    - derived_from_revenue: 266
- note: Blank cells are expected in optional columns only: Customer Phone: 22 blank; derived_from_revenue: 266 blank. 'derived_from_revenue' is a flag column that is blank unless Units/Unit Price was back-filled from Revenue; Customer Phone is blank where the raw file had no phone. Required columns have 0 blanks (enforced by the validation step).

Step 6: build workbook (2026-10-18 09:31:37 UTC)

Built sales_cleanup_before_after.xlsx: 300 raw rows -> 272 clean rows; native PivotTable + SUMIFS summary + chart

- xlsx: sales_cleanup_before_after.xlsx
- preview_png: sales_cleanup_before_after_summary_preview.png
- pivot_anchor: Summary!A19
- excel_postprocess: >Sum of Revenue<, core props, pivot captions, refreshOnLoad, tableStyleInfo
- raw_sheet: Raw (Messy) = Raw demo file (deliberately messy synthetic data)

- document_properties: title='Sales Data Cleanup - Before & After (SYNTHETIC sample)', author/creator and lastModifiedBy left empty