Customer Data Cleaning

SYNTHETIC / SIMULATED PRACTICE PROJECT

Python, CSV and Excel-compatible spreadsheet outputs. AI-assisted dataset design with deterministic cleaning and independent reconciliation.

Source rows	Unique customer IDs	Duplicates excluded	Review entries
500	470	30	77

The brief

Prepare a customer-level export for a fictional online store. The generated source includes inconsistent casing, phone layouts, dates, city spelling, repeated IDs and missing contact fields. Preserve source rows and document uncertainty.

Cleaning decisions

Normalize: trim whitespace, standardize name/email casing, format phones with +62, and map known city variants. Formatting does not verify contact ownership or deliverability.
Deduplicate: retain the first observed row per Customer_ID. Exclude later rows without adding their Total_Spent values. No duplicate spend conflicts were found in this dataset.
Defer: leave missing email/address values and the six quarantined registration dates blank; record source row, issue, treatment and next action in the review queue.

Date convention and limitation

The generator uses YYYY-MM-DD, DD-MM-YYYY and MM/DD/YYYY. The pipeline quarantines the literal 03/04/2026 marker. Other slash dates follow the known MM/DD convention, including dates that would be ambiguous without that context. This is not a general ambiguity detector. A production import must confirm the source locale.

What I validated

Reproduced all 470 cleaned CSV rows and all 77 review queue entries from the original 500 rows. Matched the original workbook clean table, review queue, 60-row raw sample, dashboard KPIs, city counts and README totals against the CSVs. Checked retained date chronology. The original workbook is a static snapshot with no formulas.

Integrity

Self-directed practice, not a client engagement or measured business improvement. “Unique customer” means distinct synthetic Customer_ID, not a verified person. Generated contact values stay out of the public evidence package.

Count the right thing.

Review category	Entries	Handling
Duplicate Customer_ID	30	Exclude later source rows; retain audit event.
Missing email	25	Keep blank; request a verified detail.
Missing address	16	Keep blank; request from the source owner.
Ambiguous marker date	6	Keep date blank; confirm date convention.
Total queue entries	77	30 + 25 + 16 + 6

The 77 entries touch 72 distinct IDs. Excluding duplicate events leaves 47 unresolved field issues across 46 retained IDs. An ID can have multiple issues. Six missing registration dates leave 464 populated dates out of 470.

Customer ID distribution

Jakarta: 89 / Malang: 82 / Surabaya: 70 / Yogyakarta: 44 / Bandung: 37 / Semarang: 37 / Bogor: 35 / Denpasar: 31 / Makassar: 28 / Medan: 17. Total: 470 IDs. These generated city counts are not real market data.

Spend definition

Total_Spent is an invented customer-level lifetime-spend field in unspecified synthetic units. The retained sum is 1,197,969,255. It is not revenue or growth. Duplicate copies are not summed.

Evidence and privacy

Public downloads contain a redacted 470-row CSV, the 77-entry review queue, verified metric JSON and this revised PDF. The redacted CSV retains IDs, city, dates, synthetic spend and missing-email/address indicators. It omits names, email addresses, phone numbers and street addresses. Generated phone numbers could resemble real contacts. The original workbook and unredacted inputs remain local audit files.

Skills demonstrated

Remote Data Entry: record consistency and missing-field checks.
Customer Operations: data hygiene and documented escalation.
Sales Support: customer-level reporting and duplicate control.
Virtual Assistant: spreadsheet organization and reviewable handoffs.

Lessons and production limits

Define the counting unit, state the date convention, and keep formatting separate from verification.
First-observed-row retention suits this generated source; production survivorship needs source priority or a verified timestamp. Names repeat across IDs and should not be used as a sole identity key.

Case study destination (deployment pending):
https://ridhohariadi.com/projects/customer-data-cleaning/