

## Station 4 (bonus) — tidy an inconsistent table

Source file: `przyklad-korpus/poznan-1956/statystyki/ofiary-i-aresztowania.csv`

The table is dirty on purpose: mixed separators (; and ,), decimal commas, dates in 4 formats, PL+EN headers, duplicates, one row with **swapped columns**, and conflicting values.

**Prompt (paste it, then paste the file contents underneath)**

Tidy the data below into one clean table. Rules:

- one column separator,
- one date format: YYYY-MM-DD,
- numbers as numbers (move qualifiers like "co najmniej", "ok." into a "note" column),
- remove duplicates,
- fix the row where the columns are swapped,
- if two sources give different values for the same thing, keep both and flag the discrepancy.

Give the result as a Markdown table.

**What to expect:** a uniform table; the model itself catches the duplicate “zabici cywile 50” and the row “650;ranni;...” with swapped columns.

**Harder variant — export to clean CSV**

The same, but give the result as valid CSV (separator: comma, decimal point, headers in Polish). Ready to paste into Excel / Google Sheets.

**Why it matters in research:** data from different sources is never consistent. This is the first step of any quantitative analysis. **Always check the numbers** — the model tidies the format, it isn't a source of truth.