

This table looks fine. It is not.

D8A.ACADEMY/RESOURCES/DATA-CLEANING-CHECKLIST

1 · STRUCTURE <table><tr><td>row count</td><td>matches the source system?</td></tr><tr><td>one row =</td><td>one entity, stated out loud</td></tr><tr><td>id unique</td><td>COUNT(*) = COUNT(DISTINCT id)</td></tr><tr><td>dtypes</td><td>numbers numeric, dates dates</td></tr></table>	row count	matches the source system?	one row =	one entity, stated out loud	id unique	COUNT(*) = COUNT(DISTINCT id)	dtypes	numbers numeric, dates dates	2 · MISSING VALUES <table><tr><td>real NULLs</td><td>count them per column</td></tr><tr><td>fake NULLs</td><td>N/A, "", "none", 0, -999</td></tr><tr><td>pattern</td><td>random, or one source/period?</td></tr><tr><td>the rule</td><td>drop, impute or flag: write it down</td></tr></table>	real NULLs	count them per column	fake NULLs	N/A, "", "none", 0, -999	pattern	random, or one source/period?	the rule	drop, impute or flag: write it down		
row count	matches the source system?																		
one row =	one entity, stated out loud																		
id unique	COUNT(*) = COUNT(DISTINCT id)																		
dtypes	numbers numeric, dates dates																		
real NULLs	count them per column																		
fake NULLs	N/A, "", "none", 0, -999																		
pattern	random, or one source/period?																		
the rule	drop, impute or flag: write it down																		
3 · DUPLICATES <table><tr><td>exact</td><td>full-row duplicates first</td></tr><tr><td>soft</td><td>TRIM + LOWER before comparing</td></tr><tr><td>business key</td><td>same email, two customer ids</td></tr><tr><td>keep rule</td><td>latest? richest? document it</td></tr></table>	exact	full-row duplicates first	soft	TRIM + LOWER before comparing	business key	same email, two customer ids	keep rule	latest? richest? document it	4 · VALUES & TYPES <table><tr><td>ranges</td><td>age 214, amount -50: possible?</td></tr><tr><td>categories</td><td>FR, fr, France = one value</td></tr><tr><td>string numbers</td><td>strip "1,204" and "\$" then cast</td></tr><tr><td>units</td><td>cents or euros? k or absolute?</td></tr><tr><td>outliers</td><td>inspect before deleting. Ever.</td></tr></table>	ranges	age 214, amount -50: possible?	categories	FR, fr, France = one value	string numbers	strip "1,204" and "\$" then cast	units	cents or euros? k or absolute?	outliers	inspect before deleting. Ever.
exact	full-row duplicates first																		
soft	TRIM + LOWER before comparing																		
business key	same email, two customer ids																		
keep rule	latest? richest? document it																		
ranges	age 214, amount -50: possible?																		
categories	FR, fr, France = one value																		
string numbers	strip "1,204" and "\$" then cast																		
units	cents or euros? k or absolute?																		
outliers	inspect before deleting. Ever.																		
5 · DATES <table><tr><td>one format</td><td>parse all to ISO before anything</td></tr><tr><td>impossible</td><td>signups in 1970 or next year</td></tr><tr><td>timezones</td><td>UTC or local? Pick and convert</td></tr><tr><td>gaps</td><td>missing days = missing loads</td></tr></table>	one format	parse all to ISO before anything	impossible	signups in 1970 or next year	timezones	UTC or local? Pick and convert	gaps	missing days = missing loads	6 · SANITY <table><tr><td>totals</td><td>reconcile one number to the source</td></tr><tr><td>eyeball 20</td><td>random rows, read them slowly</td></tr><tr><td>log it</td><td>rows in, rows out, rows dropped why</td></tr><tr><td>rerunnable</td><td>script, not hand edits in Excel</td></tr></table>	totals	reconcile one number to the source	eyeball 20	random rows, read them slowly	log it	rows in, rows out, rows dropped why	rerunnable	script, not hand edits in Excel		
one format	parse all to ISO before anything																		
impossible	signups in 1970 or next year																		
timezones	UTC or local? Pick and convert																		
gaps	missing days = missing loads																		
totals	reconcile one number to the source																		
eyeball 20	random rows, read them slowly																		
log it	rows in, rows out, rows dropped why																		
rerunnable	script, not hand edits in Excel																		

→ Cleaning is not a chore before the analysis. It usually IS the analysis.

WORTH MEMORIZING

01 The soft-duplicate catcher

Exact duplicates are easy. These are the ones that survive.

```
SELECT LOWER(TRIM(customer)) AS c,
       COUNT(*) AS n
FROM sales
GROUP BY 1
HAVING COUNT(*) > 1
ORDER BY n DESC;
```

→ Normalize first, then group. "Ada Lovelace" and "ada lovelace " are the same person.

02 Trap: the silent cast

The column loaded fine. As text.

LOOKS LIKE	ACTUALLY IS	SO
1,204	a string	SUM() returns garbage or errors
N/A	a string	never counted as missing
007	a string	joins fail against the int 7
14/03/2026	a string	sorts as text: April before March

→ One df.dtypes or one information_schema look costs ten seconds and catches all four.