Cleaning Duplicate Rows in Excel

Idea In Short

An expense report full of duplicate charges is a common headache, and Excel's one-click remove-duplicates function hides exactly which rows disappeared. A five-step formula method using CONCATENATE, sort and IF statements finds duplicates while keeping a full audit trail.

Real Data Is Never Clean

Perfect data is a myth in consulting, regardless of the organization or the source system. A recent example: downloading six months of corporate expense charges from a credit card website, roughly 34,000 dollars across flights, hotels, rental cars, printing and meals, only to find the totals did not match personal records.1 A little sorting and filtering revealed the culprit: duplicate entries scattered through the file, not consistently, just often enough to throw off the total.

The Scale Problem

Roughly 370 expense line items contained duplicates at a rate near one in five, identifiable by matching date, vendor and amount. Eyeballing a spreadsheet and deleting duplicates manually works fine at fifteen rows and fails completely at 24,000, where scale demands a repeatable, auditable process rather than ad hoc judgment calls. Excel does include a built-in remove-duplicates function under the Data tab, but it deletes rows silently, with no visible record of what disappeared, which makes the result nearly impossible to audit afterward. A formula-based approach takes a bit longer and produces a fully trackable result instead.

Step One: CONCATENATE

The CONCATENATE function combines two or more cell values into a single cell, useful whenever a unique identifier needs to be built from multiple fields.2 Combining a date

column and an amount column into one new column creates a compound identifier unique enough to flag likely duplicates, and copying the formula down applies it to every row at once.

Step Two: Sort

Sorting the file by the newly created identifier column groups any matching rows next to each other, which is the setup required for the comparison formula in the next step to work correctly.

Step Three: An IF Statement

A simple IF statement compares each row's identifier to the one directly above it, since the prior sort step has already put matching rows adjacent to each other: if the two match, return a flag value of 1; if not, return 0. Copying that formula down the column produces a clean series marking every likely duplicate.

Step Four: Paste as Values

A step new analysts frequently skip: after building the formula columns, copy them and paste as values rather than leaving live formulas in place. This locks the 1s and 0s to their correct rows, since re-sorting a column of live formulas afterward can silently break the results.

Step Five: Sort and Review

Sorting again by the duplicate flag column groups every marked row together, in this example revealing 79 likely duplicates out of roughly 370 total entries.3 Removing those rows produces a clean data set, with a full record of exactly which entries were treated as duplicates and why.

Scaling Beyond a Single Spreadsheet

The same logic extends naturally once a data set outgrows a single spreadsheet entirely. Power Query, built into modern versions of Excel, applies essentially the same concatenate-sort-flag logic through a repeatable, refreshable query rather than a one-time set of manual

formulas, which matters once the same cleanup needs to run monthly rather than once. Analysts who build the manual version first tend to pick up the automated tools faster, because they already understand exactly what the automation is doing under the hood rather than trusting a black box the way the built-in remove-duplicates function gets trusted, and wrongly so, in the first place.

Why the Manual Method Still Earns Its Keep

Excel's built-in remove-duplicates feature is not wrong to use in every context, and it is worth being clear about when it is acceptable. For a genuinely low-stakes cleanup, where nothing downstream depends on being able to show exactly which rows were removed, the one-click function saves real time. The formula method earns its extra steps specifically when the data feeds a client deliverable, a financial reconciliation or any analysis someone might later be asked to defend, since an analyst who cannot explain which rows were treated as duplicates and why has effectively lost control of the underlying number. Building the habit of choosing the right tool for the stakes involved, rather than defaulting to whichever method is fastest, is itself a small but genuine analytical skill.

Two Caveats Worth Remembering

The method is not infallible: two genuinely separate charges could share the same date and amount by pure coincidence and get flagged incorrectly, which is why adding a third field, such as vendor name, to the concatenated identifier reduces false positives meaningfully. Second, always preserve the original data untouched on its own tab, copying it to a fresh worksheet for cleanup and cutting confirmed duplicates to a separate tab rather than deleting them outright, so the process stays reversible if a mistake surfaces later.

- 1Data Cleansing
- 2Concatenation
- 3Microsoft Excel

Summary

Real data is gunky. A CONCATENATE-and-IF formula approach finds duplicate rows while

preserving a full audit trail, unlike the one-click remove-duplicates function. Keep the original tab, cut confirmed duplicates to a separate sheet, and always sanity-check the count.