

DATA CLEANING CHECKLIST

Worked examples edition

Prepare research data for analysis without losing its provenance, introducing silent errors, or hiding consequential cleaning decisions.

CHECKLISTS / BEFORE-AND-AFTER DATA / DECISION LOGS

Illustrative survey dataset | 10 pages

01 / PREPARE AND PROTECT

Before you change anything

☐ **Preserve the raw dataset**

Keep a read-only source copy with collection date, source and original row count.

☐ **Protect participant privacy**

Store identifying fields separately where appropriate; honour consent and access limits.

☐ **Create a data dictionary**

Define each variable, type, permissible range, unit, missing code and skip logic.

☐ **Define the unit of analysis**

Establish whether each row is a participant, visit, household, event or another observation.

☐ **Set up a decision log**

For each issue record the rule, evidence, action, affected records and reviewer.

☐ **Use reproducible operations**

Save scripts, formulas or steps; never silently overwrite the only copy.



GUIDING PRINCIPLE

Cleaning should improve fidelity to the source evidence, not force the dataset towards a preferred statistical result.

02 / INSPECT SYSTEMATICALLY

Structure, values and coding

☐ **Check types and imports**

Dates, numerical variables and categories may be imported as text or parsed incorrectly.

☐ **Inspect valid ranges**

Compare observations with the instrument and plausible boundaries; investigate rather than automatically delete.

☐ **Standardise equivalent labels**

Trim whitespace and normalise case after verifying labels mean the same thing.

☐ **Check duplicates against a key**

Distinguish duplicate entries from repeat visits or legitimate multiple observations.

☐ **Validate missing codes**

Separate unanswered, refused, not applicable and system missing where the analysis requires it.

☐ **Check cross-field logic**

Look for impossible combinations and skipped questions answered unexpectedly.

**PRACTICAL SAFEGUARD**

Keep a raw column alongside a cleaned coding column when transformations are complex or need auditing.

03 / MAKE DEFENSIBLE DECISIONS

Missing values, outliers and handoff

☐ **Report missingness before changing it**

Calculate missing counts and rates and examine whether missingness varies across groups.

☐ **Choose a justified strategy**

Complete-case analysis, imputation or other approaches depend on the analysis, assumptions and data.

☐ **Investigate outliers**

An extreme but valid observation is not necessarily an error.

☐ **Compare before and after**

Check row count, distributions, category frequencies and other key summaries.

☐ **Re-run validation**

Repeat range, duplicate, missing-code and consistency checks on the cleaned file.

☐ **Package the handoff**

Save raw and cleaned data, dictionary, cleaning script, decision log and quality summary.



DO NOT DO THIS

Never remove an observation just because it makes a desired result weaker or less statistically significant.

04 / WORKED EXAMPLE

Meet the messy survey data

Fictional example: one row represents one unique student survey response. The six records below are deliberately imperfect.

ID	Age	Study hrs	Mode	Score
S01	21	5	Online	78
S02	22	4	online	81
S03	220	6	Offline	73
S04	23	NA	On-line	87
S04	23	NA	On-line	87
S05	24	8	Offline	102

Known instrument rules

Age is 17-65; score is 0-100; weekly study hours must be non-negative; Mode has two allowed values: Online and Offline. The respondent ID should be unique for this one-time survey.

FIRST-PASS DIAGNOSIS

You can spot potential duplicate ID S04, implausible age 220, an out-of-range score 102, missing study hours, and inconsistent labels for Online. Each issue needs verification before action.

05 / WORKED EXAMPLE

From raw values to decisions

How to deal with errors while distinguishing a verified correction from a guess.

Issue	Investigation	Defensible action
S02: online	Compare labels with survey options.	Trim space; code as Online.
S03: age 220	Check original submission.	If 22 verified, correct to 22; else flag missing.
S04 repeated	Compare all fields, timestamps, submission IDs.	Remove only if confirmed duplicate; otherwise review.
S05: score 102	Consult scoring rule/source form.	Correct if verified; else flag as invalid or missing.
S04: hrs NA	Confirm meaning of NA.	Keep as missing; do not silently replace.

KEEP THE DISTINCTION CLEAR

A verified transcription correction restores source information. Imputing an unknown value estimates information and requires a separate methodological justification.

06 / WORKED EXAMPLE

Before and after: one record

Suppose the source form confirms that S03 was 22, not 220. That is a verifiable correction.

RAW ROW

S03 | 220 | 6 | Offline | 73

Out-of-range age

CLEANED ROW

S03 | 22 | 6 | Offline | 73

Age verified using source form

If the source cannot be checked

Do not assume 220 means 22. Mark the age as unresolved or missing according to the documented protocol, and retain the original raw value.

What changes in the log?

Record the validation rule (age between 17 and 65), participant ID, original value, the source used for verification, action, timestamp and reviewer.

RE-CHECK IMPACT

After cleaning, compare summaries and counts. Keep a record of how many corrections, exclusions and unresolved issues remain.

07 / WORKED EXAMPLE

Missingness and extreme values

A concrete decision sequence for two common sources of analytic error.

Example A | Missing study hours

One of five distinct student IDs has study hours recorded as NA (20%). First confirm NA means unanswered, rather than not applicable or a coding error. Document the missingness and examine how analysis will handle it.

Possible approaches

A descriptive summary might report the valid n alongside the mean. A regression may need a different strategy depending on sample size, missingness assumptions and the analytic model. Do not replace the missing value with the overall mean by default.

Example B | Score of 102

The instrument says score is between 0 and 100. Check the underlying scoring sheet. If a clerical error is verified, correct it. If the cause remains unknown, flag the value as invalid or missing and state the consequence for the analysis.

WHAT ABOUT A GENUINELY EXTREME SCORE?

A score of 99 is high but in range. Retaining a valid observation is often appropriate; explore robustness rather than treating an unusual result as an automatic deletion.



08 / WORKED EXAMPLE

Completed cleaning decision log

Illustrative entries only; every correction depends on source evidence.

ID / variable	Issue / evidence	Action / impact
S02 / Mode	Raw online + trailing space; same survey option.	Standardise to Online; 1 row.
S03 / Age	220 exceeds 65; source confirms 22.	Change to 22; retain raw copy; 1 row.
S04 / ID	Duplicate-looking second row; exact form, timestamp verified.	Remove 1 confirmed repeated submission.
S04 / Hours	NA, no value in source.	Keep missing; 1 of 5 unique IDs.
S05 / Score	102 invalid; no source confirmation.	Flag invalid; do not invent a replacement.

Example quality summary

Started with 6 rows; 1 verified duplicate removed, leaving 5 unique respondent IDs. One age corrected from the source; one value standardised; one study-hours value missing; one score remains invalid pending source verification.

Your cleaning decision log

Use one entry per issue or rule and keep the corresponding code or transformation steps.

Dataset, version, source, and unit of analysis

Issue and detection rule

Affected variable(s), record IDs and counts

Source evidence and investigation

Decision taken and justification

Date, reviewer, script or transformation reference

Before/after validation and remaining limitations
