AI Researchers
Embedding Responsible AI in Research

USE CASE 3, ITEM 01: DATA PREPARATION AND CONTENT ANALYSIS

Data quality and audit

A self-guided 60-minute workshop on cleaning and preparing quantitative data

FEP University of Porto, Fábio Duarte

Use AI to accelerate checks, never to invent evidence

Co-funded by the European Union

Erasmus+ KA220-HED airesearchers.eu

CC BY-SA 4.0

Data Quality and Audit

Cleaning and Preparation

Use AI to accelerate checks — *never to invent evidence*

60 minutes, quantitative data only, 17 fictional training records

Outcome: issue log + approved cleaning plan + independent GO/STOP validation

No analysis is ready without an audit trail

A dataset becomes analysis-ready only when every relevant issue, decision and transformation can be traced and checked.

1	**Issue log**	What is wrong, where it occurs, which rule fails and what remains uncertain.
2	**Approved cleaning plan**	What will change, why, with whose approval and with what statistical consequence.
3	**Validation decision**	Evidence that row counts, IDs, missingness and rules reconcile before analysis begins.

If one transformation cannot be explained, reproduced and validated, the decision remains STOP.

AI accelerates checks; people remain accountable

Stage	AI can help	A person must decide
Audit	Profile types, missingness, duplicates and rule failures	Purpose, unit of analysis, rules and whether accuracy is evidenced
Plan	Propose reversible actions and statistical consequences	Approve, reject or amend each proposed action
Implement	Generate code for the approved plan	Run securely, version files and verify outputs
Validate	Compare raw, clean, plan and log	Authorize GO or retain STOP

AI must not invent values, infer accuracy without a source, merge conflicting IDs, delete outliers automatically or hide uncertainty.

Six steps turn raw data into a defensible decision

01	Frame the purpose	5 min	Human
02	Establish the baseline	10 min	Human first
03	Audit against rules	15 min	AI + human review
04	Decide before cleaning	15 min	Human approval
05	Validate independently	10 min	AI check + human GO
06	Debrief and reuse	5 min	Human learning

Define the analysis before judging data quality

Use case: a fictional university will report completion, mean pre/post scores, learning gain among completers and satisfaction.

Write the data contract

Unit of analysis
One row = one participant registration

Critical fields
participant_id, completed, pre_score, post_score, satisfaction

Decision period
Training delivered in March 2026

Why purpose changes quality

A missing post-score is critical for learning gain
A missing satisfaction score may be structural for a non-completer
Repeated IDs can double-count participants
Completion before start is logically impossible

Participant output: one sentence stating purpose, row unit, critical fields and period.

Inspect the raw file before asking AI

Record what is present before standardising anything. The first baseline must be independent of AI.

17 × 11
rows × variables

15
distinct IDs

1
exact duplicate copy

17 fictional training records (1 / 2)

participant_id	age	country	discipline	training_hours	pre_score	post_score	completed	satisfaction	start_date	completion_date
P001	29	Portugal	Economics	6	62	78	Yes	4	2026-03-03	2026-03-20
P002	34	PT	Engineering	8	70	85	Y	5	2026-03-04	2026-03-21
P003		Spain	Social Sciences	4	55	71	yes	4	2026-03-05	2026-03-22
P004	222	France	Economics	5	66	74	Yes	3	2026-03-05	2026-03-23
P005	41	Germany	Data Science	7,5	73	88	Yes	5	2026-03-06	2026-03-24
P006	38	portugal	Engineering	-2	68	80	Yes	4	2026-03-06	2026-03-24
P007	27	Italy	Economics	60	64	69	Yes	3	2026-03-07	2026-03-25
P008	45	Spain	data science	9	81	108	Yes	5	2026-03-07	2026-03-25

17 fictional training records (2 / 2)

participant_id	age	country	discipline	training_hours	pre_score	post_score	completed	satisfaction	start_date	completion_date
P009	31	France	Social Sciences	6	59	76	Maybe	4	2026-03-08	2026-03-26
P010	36	Germany	Engineering	7	72		Yes	4	2026-03-08	2026-03-26
P011	52	Portugal	Economics	3	61		No		2026-03-09	
P012	24	Italy	Data Science	5	67	79	Yes	4	2026-03-12	2026-03-10
P013	40	Spain	Biochemistry	8	75	86	Yes	7	2026-03-10	2026-03-27
P014	33	Portugal	Engineering	6	69	82	Yes	4	2026-03-11	2026-03-28
P015	29	France	Social Sciences	5	63	77	Yes	4	2026-03-11	2026-03-29
P001	29	Portugal	Economics	6	62	78	Yes	4	2026-03-03	2026-03-20
P014	33	Portugal	Engineering	6	69	82	Yes	6	2026-03-11	2026-03-30

Prompt A asks AI to audit without editing

AI CAN AUDIT

Act as a data-quality auditor. Audit only; do not change the file.

Context
Purpose: report completion, pre/post scores, learning gain and satisfaction.
Unit: one row = one participant registration.
Follow the following rules: (*see next slide: "Rules"*)

Checks
Report shape, missing values, inferred types, exact duplicates, repeated IDs, category/range/date/cross-field failures and review-only outliers.

Evidence and output
For every finding cite participant_id, field, observed value, violated rule, classification and recommended next check. Return a baseline summary, issue log and unresolved questions. Do not infer accuracy without a source record; do not invent, delete or impute.

Expected output: a structured issue log — not a cleaned dataset.

Rules convert vague concerns into testable checks

Completeness

Expected records and critical values exist

Uniqueness

One participant appears once where required

Consistency

Dates, status and assessment fields agree

Timeliness

The extract covers the required period

Validity

Type, format, range and category rules hold

Accuracy

Values can be verified against source records

Rules

participant_id: required + unique
age: integer 18–100
training_hours: numeric ≥ 0; >40 review
pre_score, post_score: 0–100
competion_date ≥ start_date
completed: Yes / No
satisfaction: 1–5

Accuracy needs source evidence; plausibility is not enough.

Useful audits separate four kinds of finding

Class	Meaning	Examples	Response
Confirmed failure	Explicit rule broken	P004 age 222; P006 −2 hours; P012 date order	Missing / source correction / flag
Standardisation	Same meaning, different format	PT/Portugal; Yes/Y/yes; 7,5	Approved mapping + parsing
Review only	Unusual, not proven wrong	P007 has 60 training hours	Keep + flag pending source
Structural missingness	Missing by design or process	P011 non-completer has no completion fields	Document; no automatic imputation

Repeated ID ≠ exact duplicate: P001 is a duplicate copy; P014 contains two conflicting records and needs a source check.

Choose evidence-based actions before cleaning

AI Researchers
Embedding Responsible AI in Research

P004

age = 222

Set missing + flag

Range failure does not reveal the true age.

P007

training_hours = 60

Keep + review flag

Unusual is not automatically invalid.

P014

two conflicting rows

Flag both + source-check

Never pick, merge or average without evidence.

Prompt B converts findings into an approved plan

AI CAN PROPOSE

Using the approved issue log, propose a cleaning plan only; do not execute changes.

For each issue return:
1. Action: keep / standardise / set missing / exclude exact duplicate / flag / source-check
2. Exact rule and observed evidence
3. Evidence or approval still required
4. Reversible implementation
5. Expected statistical consequence
6. Approval owner

Never invent a replacement value, merge conflicting IDs or delete a review-only outlier.

Human checkpoint: approve, amend or reject every row of the plan before asking for code.

Prompt C generates code for reproducible implementation

Generate reproducible code that implements only the attached approved cleaning plan. Preserve the raw file. Create a new clean file and a change log. For every change record participant_id, field, old value, new value, rule, action and approval reference. Stop and ask when the plan is ambiguous. Do not add unapproved transformations.

Required outputs

1 **Clean dataset**

New version; raw file remains unchanged

2 **Change log**

One traceable record per transformation

3 **Reproducible code**

Same inputs produce the same outputs

Security: run code in an approved environment, inspect the diff and version every output before validation.

Analysis stays at STOP until all six gates pass

Compare raw data, clean data, the approved plan and the audit log. Mark a gate only when evidence exists.

- ○ Raw file preserved and version noted
- ○ Every row-count change reconciled
- ○ Critical fields meet rules or limitations are explicit
- ○ Repeated IDs resolved or double counting prevented
- ○ No undocumented imputation, clipping or deletion
- ○ Clean file and audit log can be reproduced

0 / 6

conditions evidenced

STOP

Resolve or document every condition

PASS, REVIEW OR FAIL — any undocumented transformation is FAIL

Prompt D validates independently and makes no edits

AI CAN COMPARE

Compare the raw file, clean file, audit log and approved plan. Do not make new edits.

Reconcile:
- row and column counts
- participant IDs and duplicate handling
- missingness before and after
- category and parsing mappings
- every rule failure and approved action
- distribution changes in critical variables

Mark each check PASS, REVIEW or FAIL and cite the supporting evidence. Any undocumented transformation is FAIL. Conclude GO only when all six validation gates are evidenced.

Best practice: use a separate chat, reviewer or model so validation does not inherit the cleaner's assumptions.

Five questions determine whether analysis can begin

01 **Can we explain every relevant issue?**

02 **Can we justify every transformation with a rule or source?**

03 **Can we reproduce the clean file from the raw file?**

04 **Can we reconcile rows, IDs, missingness and distributions?**

05 **Can another person verify the evidence independently?**

GO

Only when every answer is YES and the six gates are evidenced.

If any answer is NO

Keep STOP. Resolve the gap or state the limitation before analysis.

Audit before editing, approve before execution, validate before analysis

Appendix — sources and prompt design checklist

A complete data-quality prompt contains six elements

1. Analytical purpose
2. Unit of analysis and critical fields
3. Data dictionary, ranges and missing codes
4. Exact task and permitted actions
5. Evidence required for every finding
6. Forbidden actions and output format

Never upload confidential or personal data unless the AI environment is institutionally approved for that data.

About this resource

Toolkit reference	UC3-01, AI Readiness Toolkit, use case 3: Data preparation and content analysis
Developed by	FEP University of Porto, Fábio Duarte
Find it at	airesearchers.eu/ai-readiness-toolkit

AI Researchers: Embedding Responsible AI in Research. Erasmus+ Cooperation Partnership in Higher Education, KA220-HED-66835D2B.

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.