Data cleaning and management

POSTER · THE MODULE AT A GLANCE

Exam Mon, 7 Dec 2026 · 180 min

1 Data cleaning steps

1 Inspect the data → 2 Missing values → 3 Outliers and errors → 4 Duplicates and types → 5 Document the choices

- A **missing value** is a data point that was not recorded or is not available (in R: **NA**). Causes: loss during collection or transfer, entry errors or omissions, not-applicable answers (skip patterns). Consequences: bias, less power, errors in the analysis.

Understand first, then clean, and record what you did.

2 Which test for which question?

Difference
Comparing means: t-test (2 groups) or ANOVA (3 or more).

Relationship
Two numeric variables: correlation. Two categorical: chi-square.

Prediction
Predict Y from X: regression (linear or logistic).

The question and the measurement level determine the test.

EXAM LIKELIHOOD

Data cleaning and missing values (NA)	92%
Outliers: investigate, do not blindly remove	92%
Data inconsistencies (notation, types, units)	88%
Transformations (log transform) against extreme values	85%
Handling NA: filtering, imputation, deletion	55%
Merging datasets (merge, primary key)	30%

EXAM TIP

- About 2 multiple-choice questions: read code and output, no code writing.

COMMON MISTAKES

- ✖ Deleting rows with NA without looking at how many there are or why they are missing.
- ✖ Leaving error codes (-999, 9999) in as real values.
- ✖ Merging without checking the row count; duplicates in the key.
- ✖ Overwriting raw columns instead of creating a new variable.
- ✖ Seeing mean imputation as neutral: it reduces the spread.
- ✖ Removing outliers automatically.

HOW TO WRITE THE ANSWER

1. Term
model or concept, English term

2. Definition
definition or features, precisely

3. Case fact
concrete fact from the case

4. Conclusion
judgement, advice or consequence

KEY CONCEPTS

data cleaning · NA · filtering · imputation · deletion · data inconsistencies · factor · outliers · log transformation · merge · rbind · primary key · new variables · normalization · standardization · long and wide