3.6 Data cleaning and evidence

Name: _____ Class: _____ Date: _____ **Total: 15 marks**

KEY WORDS data science • missing value • duplicate • cleaning

Objective

Prepare data honestly and use it to answer a defined question.

You must be able to:

- identify missing, duplicate, and impossible data
- document a cleaning decision
- distinguish an evidence-based conclusion from an unsupported claim

1 Key knowledge and worked examples

■ Clean before claiming

A survey has 200 rows, including 40 blank age values. If those rows are removed, record that decision: the analysis now describes 160 responses, not every respondent.

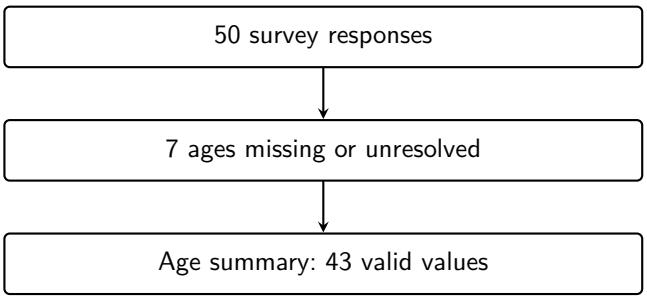
Data science answers a defined question with checked evidence. Cleaning choices change what a result can support.

■ Key knowledge: Document missingness instead of hiding it

A cleaning rule must be reproducible and linked to the analysis question. Missing values are not automatically zeros, and exclusions change the analysed population.

■ Worked example: Document missingness instead of hiding it

A survey has 50 rows: five blank ages and two age entries of 250 years. The analyst checks the two implausible entries against the original forms; if unresolved, it treats those ages as missing. An age summary then uses 43 valid ages and reports seven exclusions. A response-count chart still counts all 50 survey responses. The team does not claim that the 43 valid ages represent everyone without considering who did not supply an age.



2 Practice

2.1 State one common data-quality problem. [1]

2.2 State why a cleaning decision should be recorded. [1]

2.3 Explain why silently removing blank rows can make a conclusion less reliable. [2]

3 Exam-style questions

3.1 A dataset has duplicate records and impossible negative ages. Recommend a cleaning process and justify it. [4]

3.2 Explain why a data project must begin with a clear question rather than choosing a chart first. [4]

3.3 A file has 80 respondents, including 8 missing ages. An analyst replaces missing ages with zero and reports a mean. Explain why this is misleading and state two better reporting choices. [3]