

Exercise walk-through

3.6

Data Science

GAC Computing

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

The method

Method — 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 数据科学 answers a defined question with checked evidence. Cleaning choices change what a result can support.

2

Practice

Practice 2.1 ■ 1 mark

State one common data-quality problem.

Solution 2.1

Worked steps

- Missing values, duplicates, inconsistent spelling, or impossible entries.

B1

Practice 2.2 ■ 1 mark

State why a cleaning decision should be recorded.

Solution 2.2

Method: Clean before claiming — p.4

Worked steps

- So another person can understand how the dataset and conclusion changed.

B1

Practice 2.3 ■ 2 marks

Explain why silently removing blank rows can make a conclusion less reliable.

Solution 2.3

Method: Clean before claiming — p.4

Worked steps

- The missing rows may represent a particular group. **B1**
- Removing them can bias what the remaining data describes. **A1**

3

Exam-style

Exam-style 3.1 ■ 4 marks

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

Solution 3.1

Worked steps

- Identify and remove or merge duplicate records using a stated rule. **B1**
- Check negative ages against the source and correct, exclude, or flag them with a recorded reason. **A1**
- This removes invalid data from the analysis. **A1**
- Documenting each choice makes the conclusion auditable. **B1**

Exam-style 3.2 ■ 4 marks

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

Solution 3.2

Method: Clean before claiming — p.4

Worked steps

- A clear question determines which data and analysis are relevant. **B1**
- A chart alone may show an interesting pattern without answering a useful question. **A1**
- The project can judge whether its evidence supports the answer. **A1**
- This makes the final decision or conclusion more credible. **B1**