

3.6 Data cleaning and evidence

Name: _____ Class: _____ Date: _____ **Total: 12 marks**

KEY WORDS data science 数据科学 • missing value 缺失值 • duplicate 重复数据
• cleaning 数据清洗

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

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]

Solutions

2.1

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

2.2

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

2.3

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

3.1

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

3.2

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