Thirty Verified Explanations and Worked Answers

Notes for SHS | Topic 84 | Updated 25 November 2026 | Original educational resource

Question 1: Define dataset

Prompt: What is a dataset rather than a website?

Hint: It contains observations.

Answer: A dataset is an identifiable collection of related observations or data elements; a website can host many distinct datasets.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 2: Records and variables

Prompt: In a rectangular table, what do rows and columns usually mean?

Hint: Rows are observations.

Answer: Rows commonly represent records/observations and columns represent defined variables; this arrangement is not universal.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 3: Features and targets

Prompt: Which variable is the classification target in Iris?

Hint: Look for species labels.

Answer: The iris species/class is the target; sepal and petal measurements are features.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 4: Structured data

Prompt: Is a CSV of country-year population counts structured?

Hint: Column names are fixed.

Answer: Yes, it is tabular structured data with known fields and record meaning.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 5: Unstructured data

Prompt: Give one example of an image dataset.

Hint: Think pixel files.

Answer: An organised folder/array of handwritten digit images is an image dataset; pixel data are not ordinary spreadsheet records.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 6: CSV

Prompt: What does CSV usually store and what can go wrong?

Hint: Commas and quoting.

Answer: Delimited text tables; separators, encoding, dates and numeric identifier coercion must be checked.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 7: JSON

Prompt: Why do web APIs often use JSON?

Hint: Nested key-value records.

Answer: JSON represents structured or semistructured objects/arrays and is convenient for APIs, but may need flattening.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 8: Authoritative source

Prompt: Which site should you choose for UCI Iris?

Hint: Original host.

Answer: Use the official UCI Iris page rather than an unverifiable third-party re-upload.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 9: Metadata

Prompt: What information does a data dictionary provide?

Hint: Variable definitions.

Answer: Field meanings, units, allowed codes, source, measurement descriptions, sometimes missing-value markers.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 10: Dataset versus portal

Prompt: Is Kaggle as a whole one dataset?

Hint: It hosts many.

Answer: No. Kaggle is a platform; identify the specific dataset and original creator.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 11: Classification choice

Prompt: Choose a task from the Digits dataset.

Hint: Targets are digits.

Answer: Handwritten digit recognition is multiclass classification; output classes 0–9.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 12: Regression choice

Prompt: Which provided benchmark uses a numeric prediction outcome?

Hint: Medical benchmark with numeric target.

Answer: The scikit-learn Diabetes regression dataset; evaluation may use MAE or RMSE.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 13: Time series choice

Prompt: Which source includes daily station weather measurements?

Hint: NOAA.

Answer: NOAA GHCN Daily Summaries, with station/date, variable and unit checks.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 14: Visualization choice

Prompt: What chart suits country population by year?

Hint: Temporal order.

Answer: A labelled line chart per selected country with consistent dates and units.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 15: Download CSV safely

Prompt: What should happen before opening a CSV in Excel?

Hint: Preserve source.

Answer: Check original producer, licence, format and size, save unchanged original, then import with type settings.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 16: Missing marker

Prompt: Can a -200 air-quality sensor value be a real concentration?

Hint: Dataset codebook.

Answer: UCI Air Quality documents -200 as missing sentinel; treat according to source documentation.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 17: Train/test split

Prompt: Why keep a held-out test set?

Hint: Unseen evaluation.

Answer: To estimate performance on unseen observations under an appropriate split and prevent evaluation contamination.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 18: Documentation

Prompt: Why is metadata necessary before computing revenue?

Hint: Cancelled invoices and quantity signs.

Answer: It explains quantity, negative returns and invoice cancellation conventions, preventing invalid revenue sums.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 19: Licence interpretation

Prompt: Does a free public download prove free redistribution?

Hint: Free access is not a licence.

Answer: No. Read dataset-specific terms and original provider restrictions; if unclear, link rather than mirror.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 20: Citation

Prompt: Write a correct generic dataset citation template.

Hint: Creator, date, title.

Answer: Creator. (Year). Dataset title (Version) [Dataset]. Repository. DOI or URL; follow provider preferred citation.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 21: Outdated data

Prompt: Can 1994 Adult income data be assumed current?

Hint: Historical source.

Answer: No; report its 1994 context and discuss distribution shift before applying to modern populations.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 22: Selection bias

Prompt: Why may Portuguese school outcomes fail to represent all students?

Hint: Narrow population.

Answer: The UCI Student Performance observations come from two schools and have a limited sampling context.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 23: Sensitive fields

Prompt: What is risky about demographic predictors?

Hint: Fairness and privacy.

Answer: They may encode sensitive group distinctions and enable harmful profiling; do not use for high-stakes decisions.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 24: Leakage

Prompt: Which bike-sharing fields directly compose cnt?

Hint: casual and registered.

Answer: casual + registered = cnt; using those contemporaneous components to predict cnt leaks the target.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 25: Prediction and causation

Prompt: Does association between acidity and wine quality prove causation?

Hint: Observation versus experiment.

Answer: No. It is association in observational ratings under measurement and feature limitations.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 26: Unsafe sharing

Prompt: Should a student publish raw restricted taxi records with person-level guesses?

Hint: Privacy and permissions.

Answer: No. Avoid re-identification; publish original aggregate analysis with source link and licence checks.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 27: Portfolio choice

Prompt: Which beginner dataset gives compact numeric flower records?

Hint: Classic botanical benchmark.

Answer: Iris is a manageable first classification and visualization choice, with limits on generalization.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 28: CSV vs Parquet

Prompt: Why might NYC taxis be in Parquet?

Hint: Columnar compression.

Answer: Parquet efficiently stores large columnar tables and generally requires an engine such as pyarrow.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 29: Ethical project

Prompt: Design a safe health benchmark study.

Hint: Teaching only.

Answer: Evaluate dataset-level model errors with reproducible splits, avoid personal diagnosis and disclose benchmark limitations.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Question 30: Complete report

Prompt: Name key sections of a dataset selection report.

Hint: Provenance to limitations.

Answer: Question, provider/URL, licence, observation unit, fields, quality, methods, charts, findings, limitations, citation and reproduction steps.

Common mistake: assuming a single generic rule applies to every dataset without checking the source.

Verify third-party licences and dataset access conditions at their authoritative source before reuse. Third-party datasets are NOT included.