Ten-Task Dataset Discovery and Portfolio Capstone

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

Task 1: Choose one of the 20

Explain why its format, topic and difficulty fit your skills.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 2: Verify the primary source

Record its authoritative landing page, producer, registration and licence.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 3: Define a research question

State a testable question suitable for the observational unit.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 4: Obtain the data lawfully

Use an official download, API or library loader and note the access date.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 5: Read metadata

Describe the variables, units, period, target and source notes.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 6: Perform a quality audit

Document missing values, types, duplicates and relevant bias.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 7: Calculate summaries or a baseline model

Use appropriate descriptive statistics or a properly split model.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 8: Create three appropriate charts

Label variables, units, context and denominators.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 9: Interpret limitations

Separate correlations from causes and report unresolved problems.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

Task 10: Publish a responsible portfolio report

Include source, licence, code, charts, methods, citations and reproduction instructions.

Required evidence: source screenshot or documented URL, analytical output, interpretation, limitations.

Common mistake to avoid: assuming a download itself establishes quality, representativeness or reuse permission.

Extension: Repeat the task with a second dataset and document the differences.

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