

Exercise Sheet: Data Quality

Data, Data Storage, Data Collection

Exercise 1 – Concentration In Water Samples

A laboratory measures the concentration of a substance in water samples. The following five measurements are obtained:

2, 3, 1, 4, 20.

An automated profiling tool reports:

“Observation 20 is unusual, but its z-score is below 3.”

- (Q1) Check the affirmation of the profiling tool (Remember to consider the population standard deviation and not the sample standard deviation).

The mean is

$$\bar{x} = \frac{2 + 3 + 1 + 4 + 20}{5} = \frac{30}{5} = 6.$$

Using the sample standard deviation gives

$$s = \sqrt{\frac{4^2 + 3^2 + 5^2 + 2^2 + 14^2}{5}} = \sqrt{\frac{250}{5}} \approx 7$$

Therefore

$$z \approx \frac{20 - 6}{7} \approx 2$$

Thus 20 is not flagged by the common heuristic $|z| > 3$.

- (Q2) Does the result of the previous question provide evidence that 20 is a normal observation?

No.

It only tells us that this particular diagnostic rule does not flag the observation at the chosen threshold.

The value is still very far from the other four observations and therefore deserves investigation.

Moreover, the extreme observation affects the mean and standard deviation used to compute its own z-score.

- (Q3) Calculate Q_1 , Q_3 using the linear interpolation approach (i.e., $1 + 0.25 \times (n - 1)$ for Q_1 and $1 + 0.75 \times (n - 1)$ for Q_3). Deduce the value of the IQR. Would 20 be flagged?

The sorted observations are

1, 2, 3, 4.

Q_1 is the value at index $1 + 0.25 \times (5 - 1) = 2$, so 2, and Q_3 is the value at index $1 + 0.75 \times (5 - 1) = 3$, so 4. Therefore $IQR = Q_3 - Q_1 = 2$. The upper bound is

$$Q_3 + 1.5 \times IQR = 4 + 1.5 \times 2 = 7.$$

Since $20 > 7$, the observation would be flagged.

(Q4) The laboratory now provides additional information. Consider each case independently.

- (a) The sample containing 20 came from a polluted river immediately downstream of an industrial plant.
- (b) The laboratory technician confirms that the instrument was calibrated correctly, but the original sample label cannot be recovered.
- (c) The raw instrument log shows that the value was actually 2.0, but the decimal point was shifted during data entry.
- (d) The sample was accidentally contaminated during the laboratory procedure.

For each case, choose the most appropriate response from **remove**, **correct**, **impute**, **flag**, or **retain**, and justify your answer.

Possible answers are:

- (a) **Retain**, possibly with a flag. The value is unusual but there is contextual evidence that it may represent a real phenomenon.
- (b) **Flag** or potentially **retain**, depending on the intended analysis. There is no evidence that the measurement is wrong, but its provenance cannot be completely established.
- (c) **Correct**. The original value is known and the correction is supported by direct evidence.
- (d) **Remove**. There is evidence that the observation does not represent the phenomenon the researchers intended to measure.

The important point is that all four situations can produce the same statistical diagnosis, while requiring different responses.

(Q5) You get explained that a concentration can only vary between 0 and 10 and a choice is made to drop the value. Instead of purely removing 20, it is replaced by “unknown” and the value is flagged. The dataset becomes

2, 3, 1, 4, unknown.

What is the smallest and largest possible value of the mean? What does this tell us about what can be concluded without inventing a value for the missing observation?

The sum of the four observed concentrations is

$$2 + 3 + 1 + 4 = 10.$$

The smallest possible total is therefore $10 + 0 = 10$, and the largest possible total is $10 + 10 = 20$. Since there are five observations, the mean satisfies

$$2 \leq \bar{x} \leq 4.$$

The important point is that the missing value does not necessarily force us to invent a single estimate. We can sometimes reason about the range of possible conclusions using the information that is available.

Exercise 2 – Diagnosing A Customer Database

A marketing team maintains the following table of customer records.

id	age	country	email	last_purchase
1001	34	USA	a.dubois@mail.com	82
1002	29	United States	j.smith@mail.com	61
1003	-3	France	(empty)	95
1004	41	France	p.martin@mail.com	74
1005	52	Germany	s.klein@mail.com	15,000
1001	34	USA	a.dubois@mail.com	82

- (Q1) Identify one record or field illustrating each of: a missing value, an invalid value, an inconsistent representation, and a duplicate. For each, say whether it is checkable from the data alone or requires outside context.

- Missing: the email for id=1003.
- Invalid: age=-3 for id=1003 (violates $\text{age} \geq 0$).
- Inconsistent representation: “USA” vs. “United States” for the same country.
- Duplicate: the last row repeats id=1001 with identical values.

All four are checkable from the data alone.

- (Q2) Is last_purchase = 15,000 for id=1005 necessarily an error? What additional information would help you decide?

Not necessarily. It might be a legitimate large purchase, a data-entry error, or a unit mismatch. Deciding requires information outside the table (purchase history, currency, or a plausible range for this product line). Unlike (Q1)’s four issues, this one needs context, not just a check on the record itself.

- (Q3) Complete an audit table (field, issue, check, response, justification) for each issue found in ?? and the value from ??, assuming the intended use is “compute the average last purchase amount for a quarterly marketing report.”

Field	Issue	Check	Response	Justification
email	missing	no value	retain	irrelevant to the question
id	duplicate	key uniqueness	remove	would double-count one purchase
age	invalid (< 0)	range check	retain	irrelevant to this report
country	inconsistent	value check	correct	map to one canonical value, low-risk
last_purchase	suspicious	(needs context)	flag	no evidence yet it is wrong

- (Q4) Suppose instead the intended use is “*flag potential fraudulent transactions for manual review.*”, indicate which rows would need to change and how.

Under fraud detection, the second row’s response changes from *remove* to *flag*, and its justification changes from “double-counting” to “a repeated charge may itself indicate fraud.” The last row’s response is unchanged (*flag*), but the record now becomes the priority of the whole exercise rather than a nuisance to resolve.

Exercise 3 – Data Quality In SQL

A hospital stores information about patient admissions in a relational database. The database contains an `admission` table:

```
admission(  
    patient_id,  
    admission_date,  
    discharge_date,  
    department  
)
```

The hospital has the following rules:

1. `patient_id` must not be missing.
2. `admission_date` must not be missing.
3. If `discharge_date` is present, it cannot precede `admission_date`.
4. A patient cannot have two admissions with exactly the same admission date.
5. The department value should be within the list of the hospital’s departments: ‘Cardiology’, ‘Emergency Room’, ‘Neurology’, and ‘Surgery’.

A small extract contains:

patient_id	admission_date	discharge_date	department
1023	2026-01-04	2026-01-07	Cardiology
1041	2026-01-05	2026-01-04	Neurology
1023	2026-01-04	2026-01-07	Cardiology
1092	2026-01-08	NULL	Cardiology
NULL	2026-01-09	2026-01-12	Surgery
1138	2026-01-10	2026-01-11	Dermatology

- (Q1) Identify the records that violate the hospital’s quality rules. Classify each problem as primarily involving:

- validity;
- consistency;
- uniqueness;
- completeness.

The second record violates the consistency rule because

$$\text{discharge_date} < \text{admission_date}.$$

The third record is a duplicate of the first and therefore violates the uniqueness rule.

The fifth record has a missing `patient_id` and therefore violates the completeness requirement.

The sixth record has a department that is not in the known list. This is a validity violation.

The fourth record is not necessarily a violation: a missing discharge date may simply indicate that the admission is still ongoing.

- (Q2) Consider the first two rules, stating that the patient id and the admission date must not be missing. How can we represent this rule using SQL?

In SQL, these rule correspond to NOT NULL constraints, which can either be enforced when creating the database:

```
1 CREATE TABLE admission(  
2     patient_id int NOT NULL,  
3     admission_date DATE NOT NULL,  
4     discharge_date DATE,  
5     department varchar(255)  
6 )
```

- (Q3) Consider the rule “if a discharge date is present, it cannot precede the admission date.” Express this rule as a Boolean condition on one record. What condition would identify violations

The rule can be expressed as

$$\text{discharge_date is NULL} \vee \text{discharge_date} \geq \text{admission_date}.$$

And the violation becomes the negation of the previous condition:

$$\text{discharge_date is NOT NULL} \wedge \text{discharge_date} < \text{admission_date}$$

- (Q4) Write an SQL query that returns all records violating the rule from the previous question.

A query for rule 3 is:

```
1 SELECT *  
2 FROM admission  
3 WHERE discharge_date IS NOT NULL  
4     AND discharge_date < admission_date;
```

- (Q5) Write a query that detects duplicate (`patient_id`,`admission_date`) combinations.

One way to detect duplicate combinations is:

```
1 SELECT patient_id, admission_date, COUNT(*)  
2 FROM admission  
3 GROUP BY patient_id, admission_date  
4 HAVING COUNT(*) > 1;
```

The exact SQL syntax may vary slightly between database systems, but the important point is that the query directly implements the quality rule.

- (Q6) Suppose the full `admission` table has 3,000 records, and the query from the previous question returns 45 records involved in a duplicate (`patient_id`,`admission_date`) pair. Compute the corresponding uniqueness metric. Would you consider this rate acceptable

if the table were used (a) to generate a regulatory compliance report, or (b) to estimate average length of stay across the hospital? Justify.

The violation rate is $45/3,000 = 1.5\%$. For (a), a compliance report is typically expected to have *zero* tolerance for duplicate admissions, since each violation may itself be a reportable error. Thus, 1.5% would likely be unacceptable, and every violating record should be investigated individually. For (b), a small fraction of duplicated admissions is unlikely to noticeably shift an aggregate average computed over 3,000 records: the same 1.5% may be tolerable here.

- (Q7) The hospital discovers that transfers between departments can produce two admissions for the same patient on the same day. The database currently reports these as duplicate records. What is currently wrong in the SQL modeling?

The SQL query is correctly implementing the stated uniqueness rule, but the rule itself is too restrictive for the hospital's actual domain.

The combination

`(patient_id, admission_date)`

is not sufficient to uniquely identify an admission if a patient can have two admissions on the same day in different departments.

The quality constraint therefore needs to be reconsidered. For example, if the department distinguishes the two admissions, uniqueness might instead be required for

`(patient_id, admission_date, department)`.

Note that in SQL, uniqueness can also be enforced using a `PRIMARY KEY` constraint, provided that the chosen columns correctly identify a record.

Exercise 4 – What Exactly Did Our Cleaning Operation Do?

A dataset contains measurements of a quantity recorded by several instruments. Before analysis, the following cleaning operations are available:

- *R*: remove duplicate observations;
- *I*: replace missing values (NA) by the mean of the observed values;
- *C*: replace the sentinel value -999 by a missing value (NA).

Consider the following small dataset:

$$D = [10, 10, 20, -999].$$

- (Q1) Suppose that the final dataset is

$$[10, 20, 15].$$

A colleague tells you that the dataset was obtained from D using exactly each operation once. Determine in what order the operations were applied. Is the solution unique?

If we first apply *C* and then *R* (resp. *R* and then *C*), we obtain $[10, 20, \text{NA}]$. The mean is now

$$\frac{10 + 20}{2} = 15,$$

so $I(R(C(D))) = I(C(R(D))) = [10, 20, 15]$. In particular, we cannot re-create the cleaning history from the raw data and the clean data alone.

- (Q2) Show that R and I do not commute. Recall that two operations $f : X \rightarrow X$ and $g : X \rightarrow X$ commute if, for every $x \in X$, $f \circ g(x) = g \circ f(x)$. (*Hint: consider the result of the previous question.*)

For the dataset D above,

$$R(I(C(D))) = [10, 20, 13.33] \neq [10, 20, 15] = I(R(C(D))).$$

so R and I do not commute.

The order of the cleaning operations is therefore part of the data workflow and can affect the result of the analysis.

- (Q3) Show that R , I and C are idempotent. Recall that an operation $f : X \rightarrow X$ is idempotent if, for every $x \in X$, $f \circ f(x) = f(x)$. Does it mean that the pipeline $C \circ I \circ R$ is also idempotent?

After the first application of C , there are no occurrences of the sentinel value -999 left. Therefore applying C again makes no further change:

$$C(C(D)) = C(D).$$

Thus C is idempotent. Similarly, R and I is idempotent, since the dataset no longer contains duplicates after the application of R .

Applying the pipeline once gives

$$C \circ I \circ R(D) = [10, 20, \text{NA}].$$

Applying it again gives

$$(C \circ I \circ R)^2(D) = [10, 20, 15].$$

Thus the pipeline is not idempotent.

Idempotence is a useful property for cleaning operations because it means that applying the same cleaning step again does not change an already cleaned dataset.

- (Q4) Consider the following two datasets:

$$D_1 = [0, 0, 0, 10, 20, 30, -999]$$

$$D_2 = [0, 10, 20, 30, 30, 30, -999].$$

Apply C , then I , then R to each dataset and compare the results after each step. What information about the original data has been lost?

$$\begin{aligned}
C(D_1) &= [0, 0, 0, 10, 20, 30, \text{NA}] \\
I(C(D_1)) &= [0, 0, 0, 10, 20, 30, 10] \\
R(I(C(D_1))) &= [0, 10, 20, 30]
\end{aligned}$$

$$\begin{aligned}
C(D_2) &= [0, 10, 20, 30, 30, 30, \text{NA}] \\
I(C(D_2)) &= [0, 10, 20, 30, 30, 30, 20] \\
R(I(C(D_2))) &= [0, 10, 20, 30]
\end{aligned}$$

Both give the same output, even though the intermediate collections are different.

The multiplicities (three 0s vs three 30s), the original mean and median, and the trace of where the sentinel was and what was imputed. The pipeline is not injective, hence keep the raw data and a log.