

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

- (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?

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.
- (Q2) Is `last_purchase = 15,000` for `id=1005` necessarily an error? What additional information would help you decide?
- (Q3) Complete an audit table (field, issue, check, response, justification) for each issue found in (Q1) and the value from (Q2), assuming the intended use is “compute the average last purchase amount for a quarterly marketing report.”
- (Q4) Suppose instead the intended use is “flag potential fraudulent transactions for manual review.”, indicate which rows would need to change and how.

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.
- (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?
- (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
- (Q4) Write an SQL query that returns all records violating the rule from the previous question.
- (Q5) Write a query that detects duplicate (`patient_id`, `admission_date`) combinations.
- (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.
- (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?

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?

- (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.*)
- (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?
- (Q4) Consider the following two datasets:

$$D_1 = [0, 0, 0, 10, 20, 30, -999] \qquad D_2 = [0, 10, 20, 30, 30, 30, -999].$$

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?