

A strike that looks like an emergency

SCH07 attends at 88.2% and SCH18 at 85.2%. Build the register as a full calendar and count every missing day as an absence, and they read 66.1% and 63.9% — the worst schools in the district, on fifteen days nobody was meant to be there.

Pierre Robentz Cassion

Lesson 4 of 8

Education Programme Analysis — Enrolled is not attending

What this lesson covers

- The rows that are not there
- What happens if you do not look
- The rule
- When a closure is the finding
- The general form
- What comes next

The rows that are not there — In Python

```
import pandas as pd

attendance = pd.read_csv("school-attendance-2024.v1.csv")
roster = pd.read_csv("school-roster-2024.v1.csv")

joined = attendance.merge(roster[["student_id", "school_id"]], on="student_id")
calendar = sorted(joined["attendance_date"].unique())

days_per_school = joined.groupby("school_id")["attendance_date"].nunique()
print(f"school days in the calendar: {len(calendar)}")
print(days_per_school.sort_values().head())
```

The rows that are not there — In R

```
library(dplyr)

attendance |>
  left_join(roster, by = "student_id") |>
  summarise(days = n_distinct(attendance_date), .by = school_id) |>
  arrange(days)
```

The rows that are not there

- Twenty-two schools have all sixty days. SCH07 and SCH18 have forty-five

The rows that are not there — In Python

```
missing = {
    school: sorted(set(calendar) - set(group["attendance_date"]))
    for school, group in joined.groupby("school_id")
}
for school, dates in missing.items():
    if dates:
        print(f"{school}: {len(dates)} days, {dates[0]} to {dates[-1]}")
```

The rows that are not there — In R

Which dates, not just how many. The pattern is the diagnosis.

The rows that are not there

- Both are missing exactly 11 to 29 March — and both are missing the same fifteen days
- Consecutive missing days at multiple schools is a closure until proved otherwise — Scattered missing days at one school. . .

What happens if you do not look — In Python

```
def attendance_rate(school, missing_counts_as_absent):
    students = roster.loc[roster["school_id"] == school, "student_id"]
    marks = attendance[attendance["student_id"].isin(students)]
    present = marks["present"].isin(["true", "Y"]).sum()
    if missing_counts_as_absent:
        denominator = len(students) * len(calendar)
    else:
        denominator = marks["present"].isin(["true", "Y", "false", "N"]).sum()
    return present / denominator

for school in ("SCH07", "SCH18", "SCH01"):
    observed = attendance_rate(school, False)
    filled = attendance_rate(school, True)
    print(f"{school}: observed {observed:.1%}, grid-filled {filled:.1%}")
```

What happens if you do not look — In R

Two denominators: marks made, and student-days in the calendar.

What happens if you do not look

School	On marks made	On a full grid
SCH07	88.2%	66.1%
SCH18	85.2%	63.9%
SCH01	91.3%	91.3%

What happens if you do not look

- Twenty-two points, invented — On the grid-filled figures SCH07 and SCH18 are the two worst schools in the district by a . . .
- The unaffected school is unchanged — which is what makes the error so hard to catch: the table looks fine, most of it. . .

The rule

- Build the denominator from days the school was open, not from the calendar

The rule — In Python

```
open_days = joined.groupby("school_id")["attendance_date"].nunique()
enrolled = roster.groupby("school_id")["student_id"].nunique()
expected = (open_days * enrolled).rename("student_days_expected")

actual = joined.groupby("school_id").size().rename("marks_made")
coverage = (actual / expected).rename("mark_coverage")
print(pd.concat([expected, actual, coverage.round(3)], axis=1).head())
```

The rule — In R

Expected student-days uses each school's own open days.

When a closure is the finding — Example

Attendance, February to April 2024

Average daily attendance, all schools	88.4%	on marks made
SCH07	88.2%	
SCH18	85.2%	

SCH07 and SCH18 were closed for the fifteen school days from 11 to 29 March, a quarter of the term. Their attendance figures are computed on the 45 days they were open and are comparable with other schools; their instructional time is not.

Counting the closure days as absences would report these two schools at 66.1% and 63.9% and rank them worst in the district.

When a closure is the finding

- Report the closure as lost instructional days, separately from attendance — The two answer different questions: . . .

The general form

Course	The absent thing	What it looked like
Routine data and DHIS2 WASH analysis	A facility that did not report A monitoring round nobody drove	A district with falling coverage A district with improving functionality
Education	A day the school was closed	Two schools with a dropout emergency

The general form

- In all three, the absence is not a value and nothing in the file announces it — The habit that catches all three is the. . .

What comes next

- Attendance describes one term.

Read the full lesson, with runnable code

```
https://data-analysis.cassion.dev/courses/education-programme-analysis/  
edu-04-the-days-that-are-not-there
```