

The columns that are not there

This dataset has fourteen columns and no name, no contact detail, no free text, no incident date, no location below district and no perpetrator detail. Every one of those absences is a decision, and every one of them can be justified by what the analysis needs.

Pierre Robentz Cassion

Lesson 1 of 8

Protection and GBV Data — What not to collect

What this lesson covers

- Start with the schema, not the data
- The principle, stated as a rule about columns
- What the coarsening costs, measured
- The one that is genuinely a loss
- The 62 blank age bands
- What to write in the methodology note
- What comes next

Start with the schema, not the data — In Python

```
import pandas as pd

referrals = pd.read_csv("protection-referrals-2024.v1.csv")
print(referrals.columns.tolist())
print(f"{len(referrals):,} cases, {len(referrals.columns)} columns")
```

Start with the schema, not the data — In R

```
library(dplyr)
```

```
referrals |> glimpse()
```

Start with the schema, not the data

Not here	Why not
Name, contact details	Never leaves the case management system, under any circumstance
Free-text narrative	Identifying by construction, and no analysis reads it
Incident date	With an area and a category, frequently identifying
Location below admin2	A village plus a category is a person
Exact age	Age band answers every disaggregation question age can
Incident type, perpetrator relationship	Not shared outside the case management agency at all

Start with the schema, not the data

- Each row of that table is a decision someone made, and each one can be defended by naming the analysis it does not. . .

The principle, stated as a rule about columns

- **Collect the minimum necessary.** If no indicator needs a field, holding it is pure risk.
- **Share less than you hold.** The analysis extract is not the case file. Each step outward drops fields.
- **Never share incident-level data outside the case management agency.** Aggregate or do not send.
- **Consent governs use.** A person consenting to a referral has not consented to a research dataset.
- **Safety first, always.** Where an analysis and a survivor's safety conflict, the analysis loses.

The principle, stated as a rule about columns — In Python

```
minimum = {
    "who": ["age_band", "sex", "disability_reported"],
    "where": ["admin1", "admin2"],
    "when": ["referral_month"],
    "what": ["case_category", "service_requested"],
    "pathway": ["consent_to_refer", "referral_made", "referral_accepted",
                "days_to_first_service", "case_status"],
}
print({k: len(v) for k, v in minimum.items()})
```

The principle, stated as a rule about columns — In R

Five groups of fields, each earning its place by an indicator that needs it.

The principle, stated as a rule about columns

- Every field in this extract maps to an indicator — `days_to_first_service` exists because the 72-hour clinical standard...

What the coarsening costs, measured — In Python

```
print(referrals["age_band"].value_counts(dropna=False))  
print(referrals["referral_month"].nunique(), "distinct periods")  
print(referrals["admin2"].nunique(), "areas")
```

What the coarsening costs, measured — In R

```
referrals |> count(age_band)
referrals |> summarise(
  periods = n_distinct(referral_month),
  areas = n_distinct(admin2))
```

What the coarsening costs, measured

- **Timeliness:** measured in days from referral to service, which is already a duration and does not need a calendar. . .
- **Seasonality:** monthly is the finest resolution any protection trend is read at, and a weekly series on 1,850 cases. . .
- **Age disaggregation:** the bands are the reporting categories. A child, an adolescent, a working-age adult and an. . .
- So the coarsening costs nothing that anyone asked for — That is what makes it defensible, and it is why the argument. . .

The one that is genuinely a loss

- Dropping incident type means you cannot say which forms of violence the referral pathway serves worst — That is a real. . .

The one that is genuinely a loss — In Python

```
print("Question: does the pathway serve some incident types worse than others?")  
print("Where it can be answered: inside the case management agency")  
print("What leaves the agency: the finding, not the disaggregation")
```

The one that is genuinely a loss — In R

```
# The right answer to "we need incident type" is often "we will run it and  
# send you the result", not "no".
```

The one that is genuinely a loss

- "You cannot have the file, and here is the answer to your question" is the professional response — and it is available. . .

The 62 blank age bands — In Python

```
missing = referrals["age_band"].isna()
print(f"{missing.sum()} cases with no age band ({missing.mean():.1%}")
print(referrals.loc[missing, "case_category"].value_counts())
```

The 62 blank age bands — In R

```
referrals |> filter(is.na(age_band)) |> count(case_category)
```

What to write in the methodology note — Example

Data handling

Analysis extract holds 14 fields for 1,850 cases. It contains no name, contact detail, free text, incident date, location below admin2, exact age, incident type or perpetrator detail.

Age is recorded in five bands and time in months. Both are sufficient for every indicator reported here, and both reduce the risk that a row can be matched to a person.

Incident-level fields remain in the case management system. Analyses requiring them are run by the case management agency and reported as findings rather than shared as data.

62 cases (3.4%) have no age band. Age-disaggregated figures are computed on 1,788 cases and are labelled accordingly.

What to write in the methodology note

- Write this before the results, not in an annex — A protection report whose data handling section is at the back has. . .

What comes next

- Removing columns reduces risk and does not eliminate it.

Read the full lesson, with runnable code

```
https://data-analysis.cassion.dev/courses/protection-and-gbv-data/  
prot-01-the-columns-that-are-not-there
```