

Les colonnes qui n'y sont pas

Ce jeu de données a quatorze colonnes et aucun nom, aucune coordonnée, aucun texte libre, aucune date d'incident, aucune localisation sous le district et aucun détail sur l'auteur. Chacune de ces absences est une décision, et chacune se justifie par ce dont l'analyse a besoin.

Pierre Robentz Cassion

Leçon 1 sur 8

Données de protection et VBG — Ce qu'il ne faut pas collecter

Ce que couvre cette leçon

- Commencez par le schéma, non par les données
- Le principe, énoncé comme une règle sur des colonnes
- Ce que coûte le grossissement, mesuré
- Celle qui est une vraie perte
- Les 62 tranches d'âge vides
- Ce qu'il faut écrire dans la note méthodologique
- La suite

Commencez par le schéma, non par les données — En 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")
```

Commencez par le schéma, non par les données — En R

```
library(dplyr)
```

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

Commencez par le schéma, non par les données

Absent	Pourquoi
Nom, coordonnées	Ne sortent jamais du système de gestion de cas, en aucune circonstance
Récit en texte libre	Identifiant par construction, et aucune analyse ne le lit
Date de l'incident	Jointe à une zone et à une catégorie, fréquemment identifiante
Localisation sous l'admin2	Un village plus une catégorie, c'est une personne
Âge exact	La tranche d'âge répond à toutes les questions de désagrégation que l'âge peut traiter
Type d'incident, lien avec l'auteur	Ne sont pas partagés du tout hors de l'agence de gestion de cas

Commencez par le schéma, non par les données

- Chaque ligne de ce tableau est une décision que quelqu'un a prise, et chacune se défend en nommant l'analyse qu'elle. . .

Le principe, énoncé comme une règle sur des colonnes

- **Collecter le minimum nécessaire.** Si aucun indicateur n'a besoin d'un champ, le conserver n'est que du risque.
- **Partager moins que ce que l'on détient.** L'extrait d'analyse n'est pas le dossier. Chaque pas vers l'extérieur. . .
- **Ne jamais partager de données au niveau de l'incident hors de l'agence de gestion de cas.** Agréger ou ne pas envoyer.
- **Le consentement gouverne l'usage.** Une personne qui consent à un référencement n'a pas consenti à un jeu de données. . .
- **La sécurité d'abord, toujours.** Là où une analyse et la sécurité d'une survivante s'opposent, c'est l'analyse qui. . .

Le principe, énoncé comme une règle sur des colonnes — En 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()})
```

Le principe, énoncé comme une règle sur des colonnes — En R

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

Le principe, énoncé comme une règle sur des colonnes

- Chaque champ de cet extrait correspond à un indicateur —
`days_to_first_service` existe parce que le standard clinique...

Ce que coûte le grossissement, mesuré — En Python

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

Ce que coûte le grossissement, mesuré — En R

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

Ce que coûte le grossissement, mesuré

- **Promptitude** : mesurée en jours du référencement au service, ce qui est déjà une durée et n'exige aucune date de. . .
- **Saisonnalité** : le mois est la résolution la plus fine à laquelle se lit une tendance de protection, et une série. . .
- **Désagrégation par âge** : les tranches sont les catégories de rapportage. Un enfant, un adolescent, un adulte en âge. . .
- Le grossissement ne coûte donc rien que quiconque ait demandé — C'est ce qui le rend défendable, et c'est pourquoi. . .

Celle qui est une vraie perte

- Retirer le type d'incident empêche de dire quelles formes de violence le circuit de référencement sert le plus mal —. . .

Celle qui est une vraie perte — En 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")
```

Celle qui est une vraie perte — En R

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

- « Vous ne pouvez pas avoir le fichier, et voici la réponse à votre question » est la réponse professionnelle — et elle. . .

Les 62 tranches d'âge vides — En 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())
```

Les 62 tranches d'âge vides — En R

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

Ce qu'il faut écrire dans la note méthodologique — Exemple

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.

Ce qu'il faut écrire dans la note méthodologique

- Écrivez ceci avant les résultats, non en annexe — Un rapport de protection dont la section de traitement des données. . .

- Retirer des colonnes réduit le risque sans l'éliminer.

Lire la leçon complète, avec le code exécutable

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