

Données de protection et VBG

Analyser des données de référencement, de gestion de cas et d'incidents dans le respect des règles de sécurité et de confidentialité du secteur — y compris les analyses qu'il faut refuser de publier.

Formateur	Pierre Robentz Cassion
Niveau	Avancé
Heures apprenant	16 h
Enseignée en	Python · R
Mise à jour	28 juillet 2026
Secteurs	Protection
Normes et méthodologies	Norme humanitaire fondamentale (CHS) · Standards Sphère · Définitions d'indicateurs de l'UNICEF · Critères d'évaluation du CAD de l'OCDE

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/protection-and-gbv-data

Le contenu pédagogique est diffusé sous licence CC BY 4.0. Vous pouvez l'adapter et le rediffuser, y compris à des fins commerciales, en créditant Cassion.

Ce que vous saurez faire

- Énoncer ce qu'un jeu de données de protection ne doit pas contenir, et justifier chaque omission par ce dont l'analyse a réellement besoin
- Évaluer le risque de divulgation d'un tableau désagrégé et appliquer une règle de suppression avant qu'il ne sorte des murs
- Construire un circuit de référencement conditionné par le consentement et localiser le point où la plus grande part se perd
- Calculer la charge de dossiers par travailleur social et la relier aux mesures de qualité qui baissent quand elle monte
- Traiter un registre de cas censuré sans biaiser le délai de clôture vers le bas
- Écrire noir sur blanc qu'une courbe de dossiers mesure le signalement et non la prévalence, et en montrer la preuve

Prérequis

Aucun. Cette formation ne suppose aucune expérience préalable de la programmation.

Sommaire

I	Ce qu'il ne faut pas collecter	1
1	Les colonnes qui n'y sont pas	2
1.1	Commencez par le schéma, non par les données	2
1.2	Le principe, énoncé comme une règle sur des colonnes	2
1.3	Ce que coûte le grossissement, mesuré	3
1.4	Celle qui est une vraie perte	4
1.5	Les 62 tranches d'âge vides	4
1.6	Ce qu'il faut écrire dans la note méthodologique	4
1.7	La suite	5
2	Dix-neuf cellules à un	6
2.1	Le jeu de données est anonyme et le tableau ne l'est pas	6
2.2	Pourquoi quatre et non un	6
2.3	Une règle, écrite comme du code	7
2.4	Que faire au lieu de publier la cellule	7
2.5	Le constat d'équité survit à la règle	8
2.6	La demande qu'il faut refuser	8
2.7	La suite	9
II	Le circuit	10
3	Les 212 cas qui ne sont pas un échec	11
3.1	Le circuit compte quatre portes	11
3.2	Le consentement n'est pas une mesure de performance	12
3.3	Rapportez les trois nombres	12
3.4	Là où le consentement n'est pas la porte	12
3.5	Les contradictions à résoudre d'abord	13
3.6	La suite	13
4	L'écart s'ouvre avant que le référencement ne soit fait	14
4.1	Trouvez la porte, non le total	14
4.2	Quel service, et quelle zone	15
4.3	L'écart qui compte	15
4.4	Avant de le publier	16
4.5	Rapportez les portes, non le total	16
4.6	La suite	17
III	La charge de dossiers	18
5	Trente-trois dossiers chacun	19
5.1	La charge de dossiers est un stock, non un flux	19
5.2	Le nombre autour duquel les orientations sont écrites	20

5.3	Ce que fait une charge excessive	20
5.4	L'erreur d'effectif	21
5.5	Ce que la charge n'est pas	21
5.6	Rapportez-la comme un tableau de supervision	22
5.7	La suite	22
6	Les dossiers non clos sont les longs	23
6.1	Le calcul évident est biaisé	23
6.2	Corrigez d'abord les durées impossibles	24
6.3	Trois façons honnêtes de le rapporter	24
6.4	La censure n'est pas répartie uniformément	24
6.5	Le motif de clôture est un jugement	25
6.6	Rapportez-le en entier	26
6.7	La suite	26
IV	Ce que les nombres signifient	27
7	La zone dont les dossiers ont doublé est celle qui s'est améliorée	28
7.1	La courbe, et la lecture évidente	28
7.2	Ce qu'un compte de dossiers compte	28
7.3	Le test qui les distingue	29
7.4	Dites-le dans le rapport, dans la phrase à côté du nombre	30
7.5	La conséquence que l'on manque	30
7.6	D'où vient réellement la prévalence	30
7.7	La suite	31
8	La section qui liste ce que vous avez refusé	32
8.1	Le rapport, en entier	32
8.2	Pourquoi la section des refus est dans le document	33
8.3	Les quatre phrases qui portent le rapport	33
8.4	L'habitude que ce cours visait	34
8.5	La suite	34

À propos de cette formation

C'est le seul cours de la plateforme où la bonne réponse à une question analytique est parfois de **refuser de la produire**. Ce n'est pas une réserve accolée à la matière ; c'est la matière.

Le public est réel. Les gestionnaires de l'information de protection et VBG se voient demander chaque semaine un tableau qui identifierait une survivante, généralement par quelqu'un aux intentions parfaitement bonnes qui n'y a pas réfléchi, et la compétence enseignée consiste à reconnaître la demande, refuser le tableau, et proposer ce qui répond sûrement à la question sous-jacente.

Le cours s'appuie sur deux fichiers. Le **jeu de données de référencement** est apparu dans des cours antérieurs comme problème de jointure et de désagrégation ; il est ici analysé pour ce qu'il est, et **ce qu'il ne contient pas est la première leçon**. Il gagne un **registre de gestion de cas** — travailleur social, mois d'ouverture, mois de clôture, motif de clôture — car une analyse de circuit ne peut pas répondre aux deux questions sur lesquelles un superviseur est jugé.

Trois résultats structurent le cours, tous calculés à partir de ces fichiers.

38,8 % des cas atteignent un service et 43,8 % des cas consentants y parviennent, et l'écart entre ces dénominateurs est la décision d'une personne et non une défaillance de programme. Le consentement conditionne le circuit, et compter un cas sans consentement comme un échec fausse la performance et travestit la personne.

Les cas avec handicap signalé aboutissent à 26,7 % contre 46,2 %. Cet écart vaut dix-neuf points, c'est la raison même de désagréger, et c'est aussi le constat le plus susceptible d'être perdu dans un tableau trop petit pour être publié.

Les travailleurs sociaux d'une zone portent 33,2 dossiers ouverts chacun contre 14,8 dans une autre, et la zone surchargée tient 1,78 revue de plan par dossier contre 2,45, et perd le contact avec 38,9 % des dossiers qu'elle clôt. Trois tableaux, une seule cause.

Python et R tout du long. Vous devriez avoir suivi le module 3 et les cours antérieurs du module 4 — ce cours suppose que vous savez défendre un dénominateur, lire un registre censuré, et écrire une limite comme une décision.

Première partie

Ce qu'il ne faut pas collecter

CHAPITRE 1

Les colonnes qui n’y sont pas

Leçon 1 sur 8 · 120 min

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.

1.1 Commencez par le schéma, non par les données

```
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")

library(dplyr)

referrals |> glimpse()
```

Quatorze colonnes pour 1 850 cas de protection. Listez maintenant ce que contient un système de gestion de cas et que cet extrait ne contient pas :

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 t
Type d’incident, lien avec l’auteur	Ne sont pas partagés du tout hors de l’agence de gestion de cas

Chaque ligne de ce tableau est une décision que quelqu’un a prise, et chacune se défend en nommant l’analyse qu’elle n’empêche pas. C’est le test : non pas « est-ce sensible » mais « que serais-je incapable de calculer sans cela ».

1.2 Le principe, énoncé comme une règle sur des colonnes

Les principes de gestion de l’information VBG s’enseignent habituellement comme de l’éthique. Ce sont aussi, concrètement, des instructions sur un schéma :

- **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 retire des champs.
- **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 de recherche.
- **La sécurité d'abord, toujours.** Là où une analyse et la sécurité d'une survivante s'opposent, c'est l'analyse qui cède.

```
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()})
```

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

Chaque champ de cet extrait correspond à un indicateur. `days_to_first_service` existe parce que le standard clinique de 72 heures pour les soins de santé VBG l'exige ; `disability_reported` existe parce que l'écart d'équité est le constat que ce jeu de données a été bâti pour faire apparaître. Un champ sans indicateur derrière lui n'a rien à faire dans l'extrait.

1.3 Ce que coûte le grossissement, mesuré

L'affirmation selon laquelle le grossissement ne coûte rien doit être vérifiée plutôt qu'assé-
née.

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

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

Cinq tranches d'âge, douze mois, six zones. Demandez maintenant ce qu'ajouteraient un âge exact et une date exacte :

- **Promptitude** : mesurée en jours du référencement au service, ce qui est déjà une durée et n'exige aucune date de calendrier.
- **Saisonnalité** : le mois est la résolution la plus fine à laquelle se lit une tendance de protection, et une série hebdomadaire sur 1 850 cas serait du bruit.
- **Désagrégation par âge** : les tranches sont les catégories de rapportage. Un enfant, un adolescent, un adulte en âge de travailler et une personne âgée sont les quatre distinctions que fait tout indicateur de protection.

Le grossissement ne coûte donc rien que quiconque ait demandé. C'est ce qui le rend défendable, et c'est pourquoi l'argument est analytique plutôt que simplement prudent.

1.4 Celle qui est une vraie perte

Toutes les omissions ne sont pas gratuites, et prétendre le contraire est la façon dont un argument de protection des données perd sa crédibilité.

Retirer le type d'incident empêche de dire quelles formes de violence le circuit de référencement sert le plus mal. C'est une perte analytique réelle, et la réponse n'est pas que cela n'a pas d'importance. La réponse est que cette analyse a sa place à l'intérieur de l'agence de gestion de cas, conduite par ceux qui détiennent les données, et que ce qui en sort est la conclusion et non le fichier.

```
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 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 est disponible bien plus souvent qu'un refus sec ou qu'une remise discrète.

1.5 Les 62 tranches d'âge vides

```
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())
```

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

La tranche d'âge est à la fois la désagrégation la plus employée et le champ le plus souvent manquant dans les données d'admission réelles, parce qu'elle se demande à un moment où la demander est intrusive. **Soixante-deux cas font 3,4 % et ne sont pas répartis uniformément**, si bien qu'un tableau désagrégé par âge a un dénominateur plus petit que le tableau de circuit à côté de lui et doit le dire.

L'envie d'imputer est ici mauvaise pour une raison propre à ce secteur : une tranche d'âge imputée, rattachée à un cas réel dans un jeu de données de protection, est un attribut fabriqué d'une personne identifiable.

1.6 Ce qu'il faut écrire dans la note méthodologique

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.

Écrivez ceci avant les résultats, non en annexe. Un rapport de protection dont la section de traitement des données est à la fin a invité chaque lecteur à sauter la seule partie qui gouverne ce que le reste a le droit de dire.

1.7 La suite

Retirer des colonnes réduit le risque sans l'éliminer. La leçon suivante porte sur ce qui arrive quand on croise les colonnes restantes, et sur le point où un jeu de données parfaitement anonyme produit un tableau qui identifie quelqu'un.

CHAPITRE 2

Dix-neuf cellules à un

Leçon 2 sur 8 · 120 min

Croisez zone, catégorie, tranche d'âge et sexe et ce jeu de données anonyme produit 200 cellules, dont 85 contiennent quatre cas ou moins et 19 en contiennent exactement un. Dans un jeu de données VBG, une cellule à un est une personne, et le tableau ne peut pas sortir des murs.

2.1 Le jeu de données est anonyme et le tableau ne l'est pas

Tout identifiant direct a disparu. Cela protège contre la lecture d'une ligne permettant de savoir de qui il s'agit. Cela ne protège pas contre le **comptage**.

```
import pandas as pd

referrals = pd.read_csv("protection-referrals-2024.v1.csv")

cells = referrals.groupby(
    ["admin2", "case_category", "age_band", "sex"], dropna=False
).size()

print(f"{len(cells)} cells")
print(f"    holding 1 case: {(cells == 1).sum()}")
print(f"    holding 4 or fewer: {(cells <= 4).sum()}")

library(dplyr)

referrals |>
  count(admin2, case_category, age_band, sex) |>
  summarise(cells = n(), of_one = sum(n == 1), of_four_or_fewer = sum(n <= 4))
```

200 cellules, dont 85 avec quatre cas ou moins, et 19 avec exactement un.

Une cellule à un dit : dans ce district, il y a eu cette année un cas de VBG concernant une fille de 0 à 11 ans. Quiconque travaille dans ce district peut savoir précisément de qui il s'agit. Le jeu de données ne l'a jamais nommée ; le tableau l'a fait.

2.2 Pourquoi quatre et non un

L'instinct est de ne supprimer que les cellules à un. Cela ne suffit pas, pour deux raisons.

Une cellule à deux identifie les deux personnes l'une à l'autre, et à quiconque en connaît une.

La différenciation récupère les cellules supprimées. Si un total de ligne est publié et que toutes les cellules de cette ligne sauf une le sont aussi, la manquante est une soustraction.

```
by_area = referrals.groupby(["admin2", "case_category"]).size().unstack()
print(by_area)
print("\nIf one cell is suppressed and the row total is printed,")
print("the suppressed cell is the total minus the others.")
```

```
referrals |> count(admin2, case_category) |>
  tidyr::pivot_wider(names_from = case_category, values_from = n)
```

Une règle de suppression exige donc un seuil et une suppression secondaire : supprimer les petites cellules, puis en supprimer assez d'autres pour que les petites ne soient pas récupérables par soustraction. Un seuil de cinq est le plancher courant dans ce secteur ; certaines agences emploient dix.

2.3 Une règle, écrite comme du code

```
THRESHOLD = 5

def suppress(counts, threshold=THRESHOLD):
    """Primary suppression, then a secondary pass so rows cannot be differenced."""
    table = counts.copy()
    small = table < threshold
    table = table.mask(small, other=pd.NA)

    # Secondary: if a row has exactly one suppressed cell, the row total gives
    # it away. Suppress the next smallest cell in that row as well.
    for index, row in table.iterrows():
        if row.isna().sum() == 1:
            remaining = counts.loc[index][~small.loc[index]]
            if len(remaining):
                table.loc[index, remaining.idxmin()] = pd.NA
    return table

print(suppress(by_area))
```

```
suppress <- function(x, threshold = 5) ifelse(x < threshold, NA, x)
# Plus the secondary pass: a row with exactly one NA is not suppressed at all.
```

Écrivez la règle comme une fonction et appliquez-la à chaque tableau. Une règle appliquée à l'œil est appliquée de façon inégale, et le seul tableau où on l'oublie est celui qui sera transféré.

2.4 Que faire au lieu de publier la cellule

La suppression est le dernier recours. Trois meilleures réponses viennent avant, et chacune préserve l'information que le demandeur voulait réellement.

Agréger vers le haut. Retirez une dimension. Zone par catégorie n'a aucune cellule sous cinq ; zone par catégorie par âge par sexe en a quatre-vingt-cinq.

```
coarse = referrals.groupby(["admin2", "case_category"]).size()
print(f"cells under 5: {(coarse < 5).sum()} of {len(coarse)}")
```

```
referrals |> count(admin2, case_category) |> summarise(small = sum(n < 5))
```

Rapporter le taux, non l'effectif, sur un dénominateur assez grand pour le porter. « L'aboutissement est de 31 % pour les cas avec handicap signalé » est publiable ; « 3 cas sur 11 à Mirebalais » ne l'est pas.

Répondre à la question plutôt que fournir le tableau. Un demandeur qui réclame

zone par âge par sexe veut généralement savoir si les services atteignent les enfants. C'est un nombre unique sur un grand dénominateur, et il est publiable.

2.5 Le constat d'équité survit à la règle

La crainte est que la suppression détruise l'analyse. Vérifiez-le.

```
def completion(frame):
    consenting = frame[frame["consent_to_refer"]]
    reached = consenting[consenting["referral_accepted"] &
                          consenting["days_to_first_service"].notna()]
    return len(reached), len(consenting), len(reached) / len(consenting)

disability = referrals["disability_reported"].isin([True, "true", "Yes"])
print("reported:      %d/%d = %.1f%%" % completion(referrals[disability]))
print("not reported: %d/%d = %.1f%%" % completion(referrals[~disability]))

referrals |>
  filter(consent_to_refer) |>
  summarise(reached = sum(referral_accepted & !is.na(days_to_first_service)),
            n = n(), .by = disability_reported) |>
  mutate(completion = reached / n)
```

26,7 % contre 46,2 %, sur des dénominateurs de 202 et 1 436. L'écart d'équité de dix-neuf points — le constat le plus important de ce jeu de données — est calculé sur des dénominateurs très supérieurs à tout seuil de suppression et il est entièrement publiable.

Les analyses qui échouent au test de divulgation ne sont presque jamais celles qui comptent. Une désagrégation à quatre entrées sur 1 850 cas répond rarement à une question que quelqu'un a posée; elle résulte généralement du fait d'avoir versé toutes les variables catégorielles dans un groupby pour voir ce qui en sort.

2.6 La demande qu'il faut refuser

Certaines doivent être refusées, et il vaut la peine de répéter la phrase.

Request: case counts by commune, month and incident category.
Response: declined.

Commune-by-month cells hold zero to three cases. At that resolution a count identifies individuals to anyone working in the area, and the dataset cannot be re-identified back but a person can be.

What I can provide instead:

- the same counts at district by quarter, no cell below five
- the completion rate and time to service for each commune, on annual denominators
- a written answer to the question the request is for, if you tell me what decision it feeds

If the requirement is genuinely commune-level operational planning, the right route is a data sharing agreement with the case management agency, not an extract.

Proposez trois solutions de remplacement et demandez quelle décision la demande alimente. Un refus sans alternative est remonté puis annulé par quelqu'un qui n'a

pas lu cette leçon ; un refus avec trois alternatives se termine généralement par l'acceptation de l'une d'elles.

2.7 La suite

Les règles sont posées. L'unité suivante porte sur l'analyse elle-même — en commençant par la porte à l'entrée du circuit qui décide qui figure au dénominateur, et ce n'est pas une décision de qualité des données.

Deuxième partie

Le circuit

CHAPITRE 3

Les 212 cas qui ne sont pas un échec

Leçon 3 sur 8 · 120 min

88,5 % de ces cas ont consenti à un référencement. Les 212 qui ne l'ont pas fait sont hors du dénominateur de performance, car une personne qui refuse un référencement exerce un droit plutôt que de révéler une lacune de service — et les compter comme des échecs fait deux erreurs à la fois.

3.1 Le circuit compte quatre portes

```
import pandas as pd

referrals = pd.read_csv("protection-referrals-2024.v1.csv")

gates = {
    "cases recorded": len(referrals),
    "consented to referral": referrals["consent_to_refer"].sum(),
    "referral made": (referrals["consent_to_refer"] &
                     referrals["referral_made"]).sum(),
    "referral accepted": (referrals["referral_made"] &
                          referrals["referral_accepted"]).sum(),
    "reached a service": (referrals["referral_accepted"] &
                          referrals["days_to_first_service"].notna()).sum(),
}

for name, count in gates.items():
    print(f"{name:24} {count:>5}")
```

```
library(dplyr)

referrals |> summarise(
  cases = n(),
  consented = sum(consent_to_refer),
  made = sum(consent_to_refer & referral_made),
  accepted = sum(referral_made & referral_accepted),
  reached = sum(referral_accepted & !is.na(days_to_first_service))
)
```

Porte	Cas	Part de la porte précédente
Cas enregistrés	1 850	—
Consentement au référencement	1 638	88,5 %
Référencement effectué	1 142	69,7 %
Référencement accepté	757	66,3 %
Service atteint	717	94,7 %

C'est la cascade du premier cours du module 4, dans un autre secteur — des dénominateurs enchaînés, chaque porte mesurée face à la précédente. L'arithmétique est la même. **La première porte ne l'est pas.**

3.2 Le consentement n'est pas une mesure de performance

Les 212 cas sans consentement peuvent être comptés de trois façons, et une seule est défendable.

Comme des échecs. L'aboutissement de bout en bout devient $717 / 1\,850 = 38,8 \%$. Cela traite la décision d'une personne comme un manquement du programme.

Comme des exclusions. L'aboutissement devient $717 / 1\,638 = 43,8 \%$ sur les cas ayant consenti. C'est le chiffre de performance.

Comme un constat à part entière. 11,5 % ont refusé, et *pourquoi* est une question de conception du service qui mérite d'être posée séparément.

```
consenting = referrals["consent_to_refer"]
reached = referrals["referral_accepted"] & referrals["days_to_first_service"].notna()

print(f"on all cases:          {reached.sum() / len(referrals):.1%}")
print(f"on consenting cases: {reached.sum() / consenting.sum():.1%}")
print(f"declined:             {(~consenting).sum()} {(~consenting).mean():.1%}")

referrals |> summarise(
  all_cases = mean(referral_accepted & !is.na(days_to_first_service)),
  consenting = sum(referral_accepted & !is.na(days_to_first_service)) / sum(consent_to_
    refer)
)
```

Compter un cas sans consentement comme un échec du circuit fait deux erreurs à la fois. Cela fausse la performance de cinq points, et cela enregistre la décision autonome d'une personne comme un défaut du système qui lui a offert un choix.

Les deux comptent et la seconde davantage. Un indicateur qui traite le refus comme un échec crée une pression sur les travailleurs sociaux pour obtenir le consentement, ce qui est l'inverse du sens même du consentement éclairé.

3.3 Rapportez les trois nombres

Referral pathway, 1,850 cases

Consented to referral	88.5%	1,638	
Declined	11.5%	212	reported separately
Reached a service, of consenting	43.8%	717	
Reached a service, of all cases	38.8%		for reference only

Publiez le taux de consentement comme une ligne à part. Un taux de consentement qui baisse est un signal sur la confiance envers le service, et il est invisible si le consentement ne sert jamais que de filtre.

3.4 Là où le consentement n'est pas la porte

Deux situations où ce raisonnement ne s'applique pas, à énoncer pour que la règle ne soit pas sur-appliquée.

Les cas de protection de l'enfance concernant un jeune enfant relèvent d'une détermination de l'intérêt supérieur plutôt que du consentement de l'enfant, et le champ de consentement enregistre la décision de l'accompagnant. La logique du dénominateur est la même et le fondement éthique est différent.

Les urgences vitales donnent lieu à un référencement salvateur sans attendre un consentement au transfert de données. Ces cas figurent dans le circuit et la colonne de consentement ne les gouverne pas.

```
by_category = referrals.groupby("case_category")["consent_to_refer"].agg(
    ["mean", "size"]
)
print((by_category * [100, 1]).round(1))
```

```
referrals |> summarise(consent = mean(consent_to_refer), n = n(),
                      .by = case_category)
```

Vérifiez si le taux de consentement diffère par catégorie avant d'appliquer une règle unique à toutes. Là où il diffère, la décision de dénominateur peut devoir différer aussi, et le rapport doit dire quels cas ont été traités de quelle façon.

3.5 Les contradictions à résoudre d'abord

```
impossible = (referrals["days_to_first_service"].notna() &
              ~referrals["referral_accepted"])
no_referral = impossible & ~referrals["referral_made"]
no_consent = impossible & ~referrals["consent_to_refer"]

print(f"time to service but referral not accepted: {impossible.sum()}")
print(f"  of which no referral was made at all:    {no_referral.sum()}")
print(f"  of which there was no consent:          {no_consent.sum()}")
```

```
referrals |> filter(!is.na(days_to_first_service), !referral_accepted) |>
  count(referral_made, consent_to_refer)
```

Onze cas consignent un délai jusqu'au premier service sur un référencement qui n'a jamais été accepté. Six d'entre eux ne portent aucun référencement effectué, et un n'avait pas de consentement. Ce sont des contradictions logiques et elles doivent être résolues avant qu'un taux d'aboutissement quelconque soit crédible, car chacune est simultanément un numérateur et un non-dénominateur.

La résolution est un jugement et doit être consignée. Croire la date de service implique que les champs du circuit sont peu fiables; croire les champs du circuit implique que la date de service est une faute de saisie. **Dites ce que vous avez cru et combien de cas cela a déplacés.**

3.6 La suite

Le circuit perd des cas à chaque porte. La leçon suivante trouve la porte où la plus grande part se perd, et le groupe de personnes pour qui chaque porte est pire.

CHAPITRE 4

L'écart s'ouvre avant que le référencement ne soit fait

Leçon 4 sur 8 · 120 min

Les cas avec handicap signalé aboutissent à 26,7 % contre 46,2 %. L'écart n'est pas à la porte du service — le consentement est identique à 89 % et 88,5 % — il s'ouvre à la porte où un travailleur social décide de faire ou non le référencement.

4.1 Trouvez la porte, non le total

Un taux d'aboutissement de bout en bout de 43,8 % vous dit que le circuit fuit. Il ne dit pas où, et les quatre portes perdent des quantités très différentes.

```
import pandas as pd

referrals = pd.read_csv("protection-referrals-2024.v1.csv")

def cascade(frame):
    consented = frame[frame["consent_to_refer"]]
    made = consented[consented["referral_made"]]
    accepted = made[made["referral_accepted"]]
    reached = accepted[accepted["days_to_first_service"].notna()]
    return pd.Series({
        "n": len(frame), "consented": len(consented), "made": len(made),
        "accepted": len(accepted), "reached": len(reached),
    })

totals = cascade(referrals)
print(totals)
print("\nloss at each gate:")
print(f" consent: {1 - totals['consented'] / totals['n']:.1%}")
print(f" referral: {1 - totals['made'] / totals['consented']:.1%}")
print(f" accepted: {1 - totals['accepted'] / totals['made']:.1%}")
print(f" service: {1 - totals['reached'] / totals['accepted']:.1%}")
```

```
library(dplyr)

referrals |> summarise(
  n = n(),
  consented = sum(consent_to_refer),
  made = sum(consent_to_refer & referral_made),
  accepted = sum(referral_made & referral_accepted),
  reached = sum(referral_accepted & !is.na(days_to_first_service))
)
```

Porte	Perte à ce point
Consentement	11,5 % — une décision, non une perte
Référencement effectué	30,3 %
Référencement accepté	33,7 %
Service atteint	5,3 %

Les deux portes du milieu perdent presque tout. Une fois le référencement accepté, 94,7 % des cas atteignent le service — le bout récepteur fonctionne. La défaillance est en amont : un référencement jamais fait, et un référencement fait puis refusé.

Ce sont deux problèmes différents avec deux responsables différents. **Un taux d'aboutissement unique masque lequel vous avez**, et un programme lisant 43,8 % investirait raisonnablement dans le service qui n'est pas le goulot.

4.2 Quel service, et quelle zone

```
by_service = referrals.groupby("service_requested").apply(cascade,
                                                         include_groups=False)
by_service["completion"] = by_service["reached"] / by_service["consented"]
print((by_service["completion"] * 100).round(1).sort_values())
```

```
referrals |>
  summarise(consented = sum(consent_to_refer),
            reached = sum(referral_accepted & !is.na(days_to_first_service)),
            .by = service_requested) |>
  mutate(completion = reached / consented) |> arrange(completion)
```

Service	Aboutissement	Référencement fait	Accepté
Appui aux moyens de subsistance	20,7 %	49 %	47 %
Juridique	30,0 %	61 %	50 %
Sécurité et protection physique	33,8 %	61 %	58 %
Psychosocial	53,2 %	77 %	72 %
Santé	57,6 %	81 %	76 %

L'appui aux moyens de subsistance aboutit au tiers du taux de la santé, et il échoue également aux deux portes du milieu — la moitié des référencements ne sont jamais faits et la moitié de ceux qui le sont sont refusés. C'est la signature d'un service connu pour être saturé : les travailleurs sociaux cessent d'y référer parce qu'ils connaissent la réponse.

Les zones vont de 29,7 % à 53,8 %, un écart assez large pour agir et assez étroit pour qu'aucune zone n'aille bien.

4.3 L'écart qui compte

```
disability = referrals["disability_reported"].isin([True, "true", "Yes"])
no_disability = referrals["disability_reported"].isin([False, "false", "No"])

for label, mask in [("reported", disability), ("not reported", no_disability)]:
    row = cascade(referrals[mask])
    print(f"{label:14} n={row['n']:>5} "
          f"consent {row['consented'] / row['n']:.1%} "
          f"made {row['made'] / row['consented']:.1%} "
          f"accepted {row['accepted'] / row['made']:.1%} "
          f"→ {row['reached'] / row['consented']:.1%}")

referrals |>
  mutate(disability = disability_reported %in% c("true", "Yes")) |>
  summarise(n = n(), consented = sum(consent_to_refer),
```

```
made = sum(consent_to_refer & referral_made),
accepted = sum(referral_made & referral_accepted),
reached = sum(referral_accepted & !is.na(days_to_first_service)),
.by = disability)
```

	n	Consentement	Référencement fait	Accepté	Aboutissement
Handicap signalé	227	89,0 %	54,5 %	56,4 %	26,7 %
Non signalé	1 623	88,5 %	71,9 %	67,3 %	46,2 %

Le consentement est identique — 89,0 % contre 88,5 %. Les personnes en situation de handicap sont exactement aussi disposées à être référées.

L'écart s'ouvre aux deux portes que contrôle le *système de services*. Un référencement est fait pour 54,5 % d'entre elles contre 71,9 %, et accepté pour 56,4 % contre 67,3 %. À la porte du service, la différence s'est composée en dix-neuf points.

Localiser l'écart à une porte précise transforme une observation en action. « Les résultats sont plus mauvais pour les personnes en situation de handicap » est un constat sur lequel personne ne peut agir. « Les travailleurs sociaux font un référencement pour 55 % d'entre elles contre 72 %, et les services receveurs en acceptent 56 % contre 67 % » nomme deux conversations avec deux parties nommées.

4.4 Avant de le publier

Trois vérifications, et la première changera vos nombres.

La colonne handicap a quatre valeurs. Une zone a saisi Yes et No au lieu de true et false, si bien qu'un groupement naïf produit quatre catégories, dont deux trop petites pour être interprétées et une silencieusement supprimée par un filtre booléen.

```
print(referrals["disability_reported"].value_counts(dropna=False))
```

```
referrals |> count(disability_reported)
```

Le handicap est auto-déclaré et sous-déclaré. 227 sur 1 850 font 12,3 %, en dessous de la plupart des estimations de population. L'écart est réel ; le dénominateur est un groupe qui a déclaré, et les personnes qui ne l'ont pas fait sont comptées dans le groupe de comparaison. Cela biaise l'écart **vers le bas** — 19 points sont donc un plancher.

Vérifiez la taille des cellules avant de désagréger davantage. Handicap par zone par service est exactement le tableau à quatre entrées que la leçon précédente a refusé de publier.

4.5 Rapportez les portes, non le total

Referral pathway, 1,638 consenting cases

Referral made	69.7%
Accepted, of those made	66.3%
Reached service, of accepted	94.7%
End to end, of consenting	43.8%

The pathway fails upstream. Once a referral is accepted 94.7% of cases

reach a service; the losses are at referral-making and acceptance.

Cases reporting a disability: 26.7% complete against 46.2%. Consent is identical (89.0% and 88.5%); the gap is entirely at referral-making (54.5% against 71.9%) and acceptance (56.4% against 67.3%).

Disability is self-reported by 12.3% of cases. Non-disclosure places some people in the comparison group, so the gap is a lower bound.

4.6 La suite

Le circuit s'arrête quand un cas atteint un service. Ce qui se passe ensuite est un dossier que quelqu'un doit porter, et l'unité suivante porte sur le registre qui enregistre qui en porte combien.

Troisième partie

La charge de dossiers

CHAPITRE 5

Trente-trois dossiers chacun

Leçon 5 sur 8 · 120 min

Les orientations situent une charge active VBG autour de 25 dossiers par travailleur. Une zone est ici en moyenne à 33,2 et culmine à 51. Elle tient aussi 1,78 revue de plan par dossier contre 2,45, et perd le contact avec 38,9 % des dossiers qu'elle clôt. Trois tableaux, une seule cause.

5.1 La charge de dossiers est un stock, non un flux

Un compte de dossiers est un flux : combien de dossiers se sont ouverts ce mois-ci. Une charge de dossiers est un stock : combien sont ouverts **à l'instant**, et c'est le nombre que vit un travailleur social.

Le calculer suppose d'étendre chaque dossier sur les mois où il a été ouvert.

```
import pandas as pd

cases = pd.read_csv("protection-case-management-2024.v1.csv")
CUTOFF = 12

def months_open(row):
    opened = int(row["opened_month"][5:7])
    closed = int(row["closed_month"][5:7]) if pd.notna(row["closed_month"]) else CUTOFF
    return range(opened, max(closed, opened) + 1)

active = [
    {"caseworker_id": row["caseworker_id"], "admin2": row["admin2"], "month": month}
    for _, row in cases.iterrows()
    for month in months_open(row)
]
active = pd.DataFrame(active)

caseload = active.groupby(["caseworker_id", "month"]).size().rename("open_cases")
print(caseload.describe().round(1))

library(dplyr)

cases |>
  mutate(opened = as.integer(substr(opened_month, 6, 7)),
         closed = coalesce(as.integer(substr(closed_month, 6, 7)), 12L)) |>
  rowwise() |>
  mutate(month = list(seq(opened, max(closed, opened)))) |>
  tidyr::unnest(month) |>
  count(caseworker_id, month, name = "open_cases")
```

C'est l'expansion en personnes-temps du premier cours du module 4, appliquée à un travailleur social plutôt qu'à un patient. Un dossier ouvert de mars à août contribue à six charges mensuelles, et un registre lu ligne à ligne ne le montre jamais.

5.2 Le nombre autour duquel les orientations sont écrites

```
by_area = caseload.reset_index().merge(
    cases[["caseworker_id", "admin2"]].drop_duplicates("caseworker_id"),
    on="caseworker_id",
)
summary = by_area.groupby("admin2")["open_cases"].agg(["mean", "max"]).round(1)
print(summary.sort_values("mean", ascending=False))
```

```
# Mean and peak monthly caseload per area.
```

Zone	Charge moyenne	Pic
Port-de-Paix	33,2	51
Gonaïves	28,2	41
Saint-Marc	24,7	43
Hinche	20,7	44
Mirebalais	18,7	32
Saint-Louis-du-Nord	14,8	26

Les orientations de gestion de cas VBG situent la charge active autour de 25. Port-de-Paix est au-dessus toute l'année et culmine au double ; Saint-Louis-du-Nord est nettement en dessous.

Rapportez la moyenne et le pic. Un travailleur à 51 dossiers ouverts un mois donné dispose, ce mois-là, d'environ quatre heures de travail par dossier déplacement et documentation compris, et une moyenne annuelle de 33 le dissimule.

5.3 Ce que fait une charge excessive

La raison pour laquelle la charge a un seuil est que des choses cèdent lorsqu'il est dépassé, et ce registre permet d'en observer deux céder ensemble.

```
quality = cases.groupby("admin2").agg(
    cases=("case_id", "size"),
    reviews_per_case=("case_plan_reviews", "mean"),
)
closed = cases[cases["closure_reason"].notna()]
quality["lost_contact"] = closed.groupby("admin2")["closure_reason"].apply(
    lambda s: (s == "lost-contact").mean()
)
print(quality.round(2).sort_values("reviews_per_case"))
```

```
cases |>
  summarise(n = n(), reviews = mean(case_plan_reviews),
            lost = mean(closure_reason == "lost-contact", na.rm = TRUE),
            .by = admin2)
```

Zone	Charge	Revue par dossier	Perte de contact
Saint-Louis-du-Nord	14,8	1,74	26,0 %
Port-de-Paix	33,2	1,78	38,9 %
Mirebalais	18,7	2,10	30,8 %
Gonaïves	28,2	2,31	24,7 %
Saint-Marc	24,7	2,38	22,1 %
Hinche	20,7	2,45	20,2 %

Port-de-Paix porte la charge la plus élevée, tient parmi les moins de revues de plan, et perd le contact avec la plus grande part des dossiers qu'elle clôt. Trois tableaux différents produits par une seule cause, et le tableau de charge est celui qui nomme la cause.

Notez que Saint-Louis-du-Nord tient aussi peu de revues sur une charge faible — les revues par dossier ne sont donc pas une fonction propre de la charge, et la lecture honnête est que Port-de-Paix a un problème de charge tandis que Saint-Louis-du-Nord en a peut-être un autre. **Un motif qui colle à trois zones et pas à la quatrième reste un motif, et dire quelle zone ne colle pas fait partie de son rapportage.**

5.4 L'erreur d'effectif

```
areas_per_worker = cases.groupby("caseworker_id")["admin2"].nunique()
print(areas_per_worker[areas_per_worker > 1])
```

```
cases |> summarise(areas = n_distinct(admin2), .by = caseworker_id) |>
  filter(areas > 1)
```

Un identifiant de travailleur social apparaît sous deux zones, car un agent muté a été réenregistré plutôt que déplacé.

Une charge calculée par agent est juste. Une charge calculée par agent et par zone est fausse, car elle répartit la charge réelle d'une personne sur deux lignes et fait paraître les deux gérables. Décidez de ce que signifie l'identifiant avant de grouper dessus, et si le registre ne peut pas vous le dire, c'est une question pour le bureau et non une hypothèse pour l'analyste.

5.5 Ce que la charge n'est pas

Ce n'est pas une mesure de productivité. Un agent à 40 dossiers ouverts ne travaille pas plus dur qu'un agent à 15 ; il est plus susceptible de faire défaut à 40 personnes lentement. Employer la charge pour classer les agents inverse la raison d'être de l'indicateur.

Elle n'est pas comparable entre types de dossiers. Un dossier de protection de l'enfance avec un plan impliquant une école et un tuteur n'équivaut pas à un référencement juridique ponctuel, et le champ `case_category` de ce registre est ce qui permet de les pondérer.

```
mix = cases.groupby(["admin2", "case_category"]).size().unstack(fill_value=0)
print((mix.div(mix.sum(axis=1), axis=0) * 100).round(1))
```

```
cases |> count(admin2, case_category) |>
  mutate(share = n / sum(n), .by = admin2)
```

Vérifiez la composition des dossiers avant de comparer des charges. Une zone avec davantage de dossiers de protection de l'enfance porte plus de travail par dossier, et une comparaison brute la pénalise.

5.6 Rapportez-la comme un tableau de supervision

Caseload, 2024, 1,108 cases across 17 caseworkers				
Area	Mean	Peak	Reviews/case	Lost contact
Port-de-Paix	33.2	51	1.78	38.9%
Gonaives	28.2	41	2.31	24.7%
Saint-Marc	24.7	43	2.38	22.1%
Hinche	20.7	44	2.45	20.2%
Mirebalais	18.7	32	2.10	30.8%
Saint-Louis-du-Nord	14.8	26	1.74	26.0%
Guidance: active caseload around 25 cases per worker.				
One caseworker identifier appears in two areas after a transfer; caseload is computed per worker, not per worker per area.				
Port-de-Paix is above the guideline all year and shows the pattern that follows: fewest case plan reviews and most cases closed for lost contact.				

Mettez l'orientation dans le tableau. Une charge de 33 ne dit rien à un lecteur qui ignore ce que vaut 25, et toute la raison d'être de la ligne est la comparaison.

5.7 La suite

Les dossiers se clôturent, et le temps qu'ils prennent est l'autre nombre sur lequel un superviseur est jugé. La leçon suivante porte sur cette mesure, et sur les 42,6 % de dossiers non encore clos qui la biaiserait si on les ignorait.

CHAPITRE 6

Les dossiers non clos sont les longs

Leçon 6 sur 8 · 120 min

Le délai médian de clôture vaut quatre mois sur les 636 dossiers clos. 472 autres sont encore ouverts, dont 140 dépassent déjà quatre mois, et la protection de l'enfance est censurée à 58 % contre 32 % pour la protection générale — la même statistique signifie donc des choses différentes selon la colonne.

6.1 Le calcul évident est biaisé

```
import pandas as pd

cases = pd.read_csv("protection-case-management-2024.v1.csv")

opened = cases["opened_month"].str[5:].astype(int)
closed = cases["closed_month"].str[5:].astype("Int64")
duration = closed - opened

print(f"closed cases: {duration.notna().sum()}")
print(f"median months to closure: {duration.median()}")
print(f"still open: {duration.isna().sum()} ({duration.isna().mean():.1%})")

library(dplyr)

cases |>
  mutate(opened = as.integer(substr(opened_month, 6, 7)),
         closed = as.integer(substr(closed_month, 6, 7)),
         duration = closed - opened) |>
  summarise(closed = sum(!is.na(duration)),
            median = median(duration, na.rm = TRUE),
            open = sum(is.na(duration)))
```

Quatre mois, sur 636 dossiers sur 1 108. Les 472 autres sont encore ouverts à la clôture de décembre, et ils ne sont pas manquants — ils sont **censurés**.

Le biais a un sens et il se voit directement.

```
still_open = 12 - opened[duration.isna()] + 1
print(f"open cases already past 4 months: {(still_open > 4).sum()}")
print(f"longest open: {still_open.max()} months and counting")

cases |> filter(is.na(closed_month)) |>
  mutate(so_far = 12 - as.integer(substr(opened_month, 6, 7)) + 1) |>
  summarise(past_four = sum(so_far > 4), longest = max(so_far))
```

140 des dossiers ouverts le sont déjà depuis plus longtemps que le dossier clos médian, et l'un l'est depuis onze mois. Chacun d'eux se clôturera à une durée supérieure à quatre mois, ou jamais. Les supprimer rend la réponse plus petite que la vérité, systématiquement.

6.2 Corrigez d'abord les durées impossibles

```
impossible = duration.notna() & (duration <= 0)
print(f"cases closing on or before the month they opened: {impossible.sum()}")
print(cases.loc[impossible, ["opened_month", "closed_month"]])
```

```
cases |> filter(!is.na(closed_month),
              closed_month <= opened_month) |> nrow()
```

Sept dossiers se clôturent un mois antérieur à leur ouverture. La contradiction est invisible dans chaque colonne prise seule et produit des durées négatives qu'un `median()` absorbe sans broncher.

Décidez et consignez : les exclure, ou traiter le mois de clôture comme une faute de saisie et le fixer au mois d'ouverture. **Sept dossiers ne déplaceront pas la médiane ; c'est la discipline de les trouver et de les déclarer qui manquera la prochaine fois si vous la sautez.**

6.3 Trois façons honnêtes de le rapporter

Rapporter la médiane avec la censure énoncée. L'option correcte la moins coûteuse.

```
Median time to closure: 4 months (636 closed cases).
472 cases (42.6%) were still open at the cut-off and are excluded; 140 of
them have already been open longer than 4 months, so the true median is
higher.
```

Rapporter une courbe d'aboutissement par mois. Quelle part des dossiers ouverts au mois m était close dans les k mois, calculée uniquement sur les dossiers disposant d'au moins k mois de suivi. C'est l'approche par cohorte et elle emploie correctement les dossiers censurés tant qu'ils ont été observés.

```
def closed_within(k):
    eligible = cases[opened <= 12 - k]
    dur = (eligible["closed_month"].str[5:].astype("Int64")
           - eligible["opened_month"].str[5:].astype(int))
    return (dur <= k).sum() / len(eligible), len(eligible)

for k in (3, 6, 9):
    share, n = closed_within(k)
    print(f"closed within {k} months: {share:.1%} (n={n})")
```

```
# For each k, restrict to cases with k months of possible follow-up.
```

Rapporter le délai de clôture pour une seule cohorte à suivi complet. Les dossiers ouverts en janvier ont onze mois d'observation ; ceux de novembre en ont un. Se restreindre à la cohorte précoce répond honnêtement à une question plus étroite.

Les trois sont défendables et la première ne suffit pas seule. Une médiane sans énoncé de censure à côté est la version qui sera citée.

6.4 La censure n'est pas répartie uniformément

```
by_category = pd.DataFrame({
    "closed": duration.notna().groupby(cases["case_category"]).sum(),
    "open": duration.isna().groupby(cases["case_category"]).sum(),
})
by_category["censored"] = by_category["open"] / by_category.sum(axis=1)
print(by_category.round(3))

cases |> summarise(closed = sum(!is.na(closed_month)),
                  open = sum(is.na(closed_month)), .by = case_category) |>
mutate(censored = open / (open + closed))
```

Catégorie	Clos	Ouverts	Censure	Médiane des clos
Protection de l'enfance	131	179	58 %	5 mois
VBG	255	173	40 %	4 mois
Protection générale	250	120	32 %	4 mois

La protection de l'enfance est censurée à 58 % et la protection générale à 32 %. Les deux médianes de ce tableau sont donc calculées sur des fractions très différentes de leurs cohortes, et l'écart entre cinq mois et quatre mois sous-estime l'écart réel.

Une censure inégale rend une comparaison de médianes trompeuse dans un sens précis : le groupe le plus censuré est celui dont la médiane est la plus sous-estimée, c'est-à-dire le groupe qui paraît déjà le plus mal loti.

6.5 Le motif de clôture est un jugement

```
print(cases["closure_reason"].value_counts(normalize=True).round(3))

cases |> filter(!is.na(closure_reason)) |> count(closure_reason) |>
mutate(share = n / sum(n))
```

Motif	Part des clôtures
Objectifs du plan atteints	29,1 %
Perte de contact	27,2 %
Retrait de la survivante	13,1 %
Clôture administrative	11,3 %
Déménagement	9,7 %
Transfert à une autre agence	9,6 %

Seules 29,1 % des clôtures correspondent à un plan mené à terme. C'est le chiffre principal, et c'est celui par lequel commence une conversation de supervision.

Mais regardez comment les motifs se répartissent par zone avant d'y croire.

```
by_area = pd.crosstab(cases["admin2"], cases["closure_reason"], normalize="index")
print((by_area[["closed-administratively", "lost-contact"]] * 100).round(1))

cases |> filter(!is.na(closure_reason)) |>
count(admin2, closure_reason) |> mutate(share = n / sum(n), .by = admin2)
```

Hinche classe 36,2 % de ses clôtures en closed-administratively contre 3,8 % à 9,8 % partout ailleurs, et son taux de perte de contact affiche 20,2 % contre jusqu'à 38,9 %. Le fourre-tout absorbe le motif dont un superviseur a besoin.

Un motif de clôture est le jugement d'un travailleur social, non une observation. Une distribution qui diffère fortement d'un bureau à l'autre est donc une question sur les bureaux avant d'être une question sur les dossiers — et le remède est une conversation sur le codage, non un ajustement statistique.

6.6 Rapportez-le en entier

Case closure, 1,108 cases

Closed by the cut-off	636	57.4%	
Still open	472	42.6%	censored, not missing
Median months to closure	4		of closed cases only
open cases already past 4 months	140		so the true median is higher

Censoring by category: child protection 58%, GBV 40%, general 32%.

Medians are not comparable across those columns.

Closure reasons (of 636 closures)

Case plan objectives met	29.1%	
Lost contact	27.2%	
Survivor withdrew	13.1%	
Closed administratively	11.3%	36.2% in Hinche, 3.8-9.8% elsewhere
Relocated	9.7%	
Transferred	9.6%	

7 cases record a closing month before their opening month and are excluded.

6.7 La suite

Tout dans cette unité porte sur des dossiers déjà dans le système. La dernière unité demande ce que signifie le nombre de dossiers lui-même — et pourquoi une zone dont le compte de dossiers a doublé est généralement celle où quelque chose s'est bien passé.

Quatrième partie

Ce que les nombres signifient

CHAPITRE 7

La zone dont les dossiers ont doublé est celle qui s'est améliorée

Leçon 7 sur 8 · 120 min

Le flux mensuel de Saint-Louis-du-Nord passe de 5,8 dossiers à 10,2 après juillet. Rien de la violence dans cette zone n'a changé. Un service a ouvert, et une courbe de dossiers mesure si les gens peuvent atteindre un bureau — jamais combien en avaient besoin.

7.1 La courbe, et la lecture évidente

```
import pandas as pd

cases = pd.read_csv("protection-case-management-2024.v1.csv")
monthly = cases.groupby(["admin2", "opened_month"]).size().unstack(fill_value=0)
print(monthly)

library(dplyr)

cases |> count(admin2, opened_month) |>
  tidyr::pivot_wider(names_from = opened_month, values_from = n, values_fill = 0)
```

La ligne d'une zone monte et reste haute.

```
area = cases[cases["admin2"] == "Saint-Louis-du-Nord"]
month = area["opened_month"].str[5:].astype(int)
before = (month <= 6).sum() / 6
after = (month >= 7).sum() / 6
print(f"Jan-Jun: {before:.1f} cases a month")
print(f"Jul-Dec: {after:.1f} cases a month  ({after / before - 1:+.0%})")

cases |>
  filter(admin2 == "Saint-Louis-du-Nord") |>
  mutate(half = if_else(as.integer(substr(opened_month, 6, 7)) <= 6, "H1", "H2")) |>
  count(half) |> mutate(per_month = n / 6)
```

5,8 dossiers par mois avant juillet, 10,2 après — une hausse de 74 %, tandis qu'aucune autre zone ne bouge de plus de onze points dans un sens ou dans l'autre.

La lecture évidente est que les incidents de protection à Saint-Louis-du-Nord ont augmenté de 74 % au second semestre. **Cette lecture est fausse, et c'est le détournement le plus courant des données de protection.**

7.2 Ce qu'un compte de dossiers compte

Un dossier existe quand une personne a atteint un service, a bien voulu se confier, et qu'un travailleur social a ouvert un fichier. Le compte est donc le produit de quatre choses, dont seule la dernière est ce que les gens y lisent :

Le compte dépend de	Ce qui change quand
L'existence d'un service à portée	L'un ouvre, l'un ferme, une route est coupée
Le fait que les gens le sachent	Une campagne de sensibilisation a lieu
La disposition à se confier	La confiance monte ou baisse ; une mauvaise expérience circule
Le fait que quelqu'un ouvre un dossier	Effectifs, charge de travail, pratique de référencement
L'incidence dans la population	—

Les quatre premières ont bougé à Saint-Louis-du-Nord en juillet, parce qu'un service a ouvert. La cinquième n'est pas mesurée par ce fichier et ne peut pas l'être.

```
print("Case count = incidence x reporting rate x service availability x recording")
print("A change in the product does not tell you which factor moved.")
```

```
# One equation, five unknowns, one observation.
```

7.3 Le test qui les distingue

Vous ne pouvez pas prouver à partir de la seule courbe qu'il s'agit d'un effet de signalement. Vous pouvez rassembler des indices, et trois vérifications sont généralement disponibles.

Quelque chose a-t-il changé dans le service ? L'indice le plus fort et le moins statistique. Un nouveau service, un nouvel agent de proximité, un formulaire d'admission modifié ou un partenaire qui ferme produisent tous une marche d'escalier avec une date.

La composition a-t-elle changé ? La part de protection générale à Saint-Louis-du-Nord passe de 31,4 % à 41,0 % tandis que la VBG baisse de 40,0 % à 32,8 % — les nouveaux arrivants penchent vers la catégorie la moins stigmatisée, ce à quoi ressemble un accès amélioré plutôt qu'une flambée de violence.

```
def mix(frame):
    return frame["case_category"].value_counts(normalize=True).round(3)

print("before July:", mix(area[month <= 6]).to_dict())
print("after July: ", mix(area[month >= 7]).to_dict())
```

```
cases |> filter(admin2 == "Saint-Louis-du-Nord") |>
  mutate(half = if_else(as.integer(substr(opened_month, 6, 7)) <= 6, "H1", "H2")) |>
  count(half, case_category) |> mutate(share = n / sum(n), .by = half)
```

Les autres zones ont-elles bougé ? Une hausse confinée à une zone avec une explication administrative est un effet de signalement. Une hausse dans toutes les zones au même moment a plus de chances d'être réelle, ou de venir d'un changement du système de rapportage.

```
halves = cases.assign(half=(cases["opened_month"].str[5:].astype(int) > 6))
shift = halves.groupby(["admin2", "half"]).size().unstack()
shift["change"] = shift[True] / shift[False] - 1
print((shift["change"] * 100).round(1).sort_values())
```

```
cases |> mutate(half = as.integer(substr(opened_month, 6, 7)) > 6) |>
  count(admin2, half) |> tidyr::pivot_wider(names_from = half, values_from = n)
```

Une zone à **+74,3 %** ; les cinq autres entre **−9,4 %** et **+11,3 %**. Une seule zone bougeant sept fois plus que la plus large des autres a une cause administrative, et le travail est de la trouver plutôt que de publier une affirmation sur l'incidence.

7.4 Dites-le dans le rapport, dans la phrase à côté du nombre

Cases opened, Saint-Louis-du-Nord

Jan-Jun	35 cases	5.8 per month	
Jul-Dec	61 cases	10.2 per month	+74%

A case management service opened in the area in July. This figure measures the number of people who reached a service and disclosed; it is not a measure of the number of incidents. The rise is consistent with improved access and cannot be read as a rise in violence.

The other five areas moved between -9% and +11% over the same period.

L'avertissement va à côté du nombre, non en note de bas de page. Une courbe de dossiers de protection sans cette phrase sera citée comme une tendance d'incidence dans la semaine, généralement par quelqu'un qui cite votre rapport fidèlement.

7.5 La conséquence que l'on manque

Si un compte de dossiers mesure l'accès, alors **un programme qui réussit devrait voir son compte de dossiers monter**, et un compte qui baisse est au mieux ambigu.

```
outcomes = pd.DataFrame({
    "observation": ["cases rise", "cases fall"],
    "good reading": ["outreach and access improved",
                    "incidents fell"],
    "bad reading": ["violence increased",
                   "service became unreachable, or trust was lost"],
})
print(outcomes)
```

The same number supports opposite conclusions. The context decides.

Un programme de protection ne peut donc pas employer le volume de dossiers comme indicateur de performance dans un sens ni dans l'autre, et les indicateurs qui survivent sont ceux que ce cours a construits : aboutissement du circuit, délai d'accès au service, charge de dossiers, motifs de clôture. Chacun a un dénominateur qui n'est pas la population.

7.6 D'où vient réellement la prévalence

Énoncé clairement, car la question sera posée.

Des enquêtes de prévalence en population, conçues pour cela, avec une formation spécialisée des enquêteurs et des protocoles éthiques qu'un jeu de données de service n'a pas. Elles sont coûteuses, rares, et sont le seul instrument qui réponde à la question.

Jamais de données de service, à quelque niveau de sophistication que ce soit. Aucun ajustement du taux de signalement ne sauve un dénominateur qui vaut « les personnes ayant

atteint un service ».

```
print("Question: how many women in this district experienced violence this year?")
print("Answerable from this dataset: no")
print("Instrument that answers it: a population-based prevalence survey")
```

```
# Saying "we cannot answer that with this" is the answer.
```

7.7 La suite

La dernière leçon assemble le tout en un rapport de protection — y compris la section qui énonce quelles analyses ont été demandées et refusées, et pourquoi cette section a sa place dans le document plutôt que dans un courriel que personne ne conserve.

CHAPITRE 8

La section qui liste ce que vous avez refusé

Leçon 8 sur 8 · 120 min

Un rapport de protection a besoin d'un bloc nommant les analyses demandées puis refusées, et ce qui a été proposé à la place. Sa place est dans le document plutôt que dans un courriel, car la demande sera refaite par quelqu'un qui n'a pas lu le courriel.

8.1 Le rapport, en entier

Protection and GBV data, 2024

Three admin1 areas, six admin2 areas

DATA HANDLING

read this first

Analysis extract: 14 fields, 1,850 cases. No name, contact detail, free text, incident date, location below admin2, exact age, incident type or perpetrator detail. Age in five bands, time in months.

Suppression rule: no cell below 5, with secondary suppression so rows cannot be differenced.

Incident-level analyses remain with the case management agency and are reported as findings, not shared as data.

REFERRAL PATHWAY

1,638 consenting cases

Consented to referral	88.5%	1,638 of 1,850
Declined	11.5%	212 a decision, not a failure
Referral made	69.7%	
Accepted, of those made	66.3%	
Reached service, of accepted	94.7%	
End to end, of consenting	43.8%	

The pathway fails upstream. Once accepted, 94.7% reach a service.

Livelihood support completes at 20.7% against health at 57.6%.

EQUITY

Cases reporting a disability 26.7% complete, against 46.2%
Consent is identical (89.0% and 88.5%). The gap is entirely at referral-making (54.5% against 71.9%) and acceptance (56.4% against 67.3%).
Disability is self-reported by 12.3% of cases, so the gap is a lower bound.

CASE MANAGEMENT

1,108 cases, 17 caseworkers

Mean caseload, highest area	33.2	peak 51; guidance is about 25
Mean caseload, lowest area	14.8	
Case plan reviews per case, highest-caseload area	1.78	against 2.45
Closed for lost contact, highest-caseload area	38.9%	against 20.2%
Closed by the cut-off	636	57.4%
Still open	472	42.6% censored, not missing
Median months to closure	4	of closed cases only; 140 open cases already exceed it
Case plan objectives met	29.1%	of closures

WHAT THESE NUMBERS ARE NOT

Case counts measure access and reporting, not incidence. Saint-Louis-du-Nord

opened a service in July and its monthly intake rose 74%; no conclusion about violence in that area follows.

Prevalence is not estimable from service data at any level of sophistication. It requires a population-based survey.

REQUESTED AND DECLINED

Case counts by commune, month and incident category -- declined; cells of 0 to 3 identify individuals. Offered: district by quarter, and commune-level completion rates on annual denominators.

Full case list with age and area for a partner's targeting exercise -- declined; the extract is not a beneficiary list and consent does not cover it. Offered: aggregate caseload by area to inform their staffing.

DATA QUALITY

62 cases (3.4%) have no age band; age-disaggregated figures use 1,788.

11 cases record a service time on an unaccepted referral; excluded, and the completion rate is computed without them.

7 cases close before they open; excluded from duration statistics.

One caseworker identifier spans two areas after a transfer; caseload is computed per worker.

One area records disability as Yes/No rather than true/false; normalised.

8.2 Pourquoi la section des refus est dans le document

Parce que la demande sera refaite. Par un nouveau gestionnaire de l'information, par le consultant d'un bailleur, par un collègue qui n'était pas en copie. Un refus qui vit dans une boîte mail doit être réargumenté chaque fois, généralement par quelqu'un de junior, sous pression de temps, face à quelqu'un de senior.

Parce qu'elle montre que le refus était raisonné et non réflexe. Un rapport qui liste deux demandes refusées à côté de ce qui a été proposé à la place se lit comme un jugement professionnel. Un rapport sans une telle section, venant d'un analyste qui dit non en réunion, se lit comme de l'obstruction.

Parce que c'est un cahier des charges. Deux ans de demandes refusées constituent une spécification de ce que le système d'information devrait pouvoir produire sans risque, et personne ne rassemble cela si ce n'est pas écrit quelque part de durable.

```
declined = [
  {"request": "counts by commune, month, incident category",
   "reason": "cells of 0-3 identify individuals",
   "offered": "district by quarter; commune completion rates"},
  {"request": "full case list with age and area",
   "reason": "not a beneficiary list; consent does not cover it",
   "offered": "aggregate caseload by area"},
]
```

Keep the log as data. It is the only way it survives a staff change.

8.3 Les quatre phrases qui portent le rapport

Tout ce qui précède est de la preuve. Un lecteur en retient quatre choses, et elles doivent être écrites avant les tableaux plutôt qu'extraites d'eux.

Le circuit échoue avant la porte du service. 94,7 % des référencements acceptés atteignent un service; les pertes sont à la décision de référer et à l'acceptation, qui sont deux

conversations avec deux parties nommées.

Les personnes en situation de handicap sont mises en échec par le système, non par leurs propres choix. Consentement identique, dix-neuf points d'écart d'aboutissement, entièrement aux deux portes que contrôle le système de services.

Une zone porte un tiers de dossiers de plus que l'orientation et cela se voit dans deux autres tableaux. La charge est la cause, les revues par dossier et la perte de contact sont les symptômes.

Aucun nombre ici ne mesure la quantité de violence survenue. Chaque dénominateur de ce rapport est constitué de personnes ayant atteint un service.

8.4 L'habitude que ce cours visait

Trois questions, dans l'ordre habituel de la plateforme, avec un ajout propre à ce secteur.

1. **Qu'ai-je compté, sur quoi ?** Cas consentants, référencements acceptés, dossiers ouverts, clôtures — quatre dénominateurs et aucun n'est la population.
2. **Qu'est-ce qui pourrait aussi l'expliquer ?** L'ouverture d'un service, une habitude de codage, le jugement d'un travailleur social sur ce que signifie « clôturé administrativement ».
3. **Que devrait-on faire différemment à cause de cela ?** Deux conversations sur la décision de référer, une sur l'effectif à Port-de-Paix, une sur le codage des clôtures à Hinche.
4. **Qui pourrait être exposé si ceci était publié tel quel ?** La question que les trois autres ne posent pas, et celle qui gouverne si la réponse aux trois premières sort un jour des murs.

La quatrième question est tout ce cours, et c'est la seule de cette plateforme où la bonne réponse consiste parfois à ne rien produire.

8.5 La suite

Le module 4 n'a plus qu'un cours. L'assiduité scolaire et les résultats d'apprentissage apportent un nouveau secteur et, pour la première fois dans ce module, un jeu de données que la plateforme n'a pas encore — ce qui est un problème de contenu avant d'être un problème analytique.

À propos de cette édition

Ce PDF est produit à partir de la même source que la version web de la formation ; les deux ne peuvent donc pas diverger. L'édition en ligne comporte le code exécutable, les jeux de données téléchargeables et les corrections publiées depuis la production de ce fichier.

Tous les jeux de données cités dans cette formation sont synthétiques ou entièrement anonymisés. Aucun enregistrement ne renvoie à une personne réelle.

Cette formation est publiée en anglais et en français. L'édition française est une adaptation professionnelle reprenant la terminologie employée par l'UNICEF, l'OMS, le PAM, OCHA et les ministères nationaux — et non une traduction littérale.

Consulter et exécuter la formation en ligne sur data-analysis.cassion.dev/fr/cours/protection-and-gbv-data

Le contenu pédagogique est diffusé sous licence CC BY 4.0. Vous pouvez l'adapter et le rediffuser, y compris à des fins commerciales, en créditant Cassion.

Généré le 28 juillet 2026.