

Cinq dimensions, cinq mesures

« Les données sont mauvaises » n'est pas un constat. Exactitude, complétude, actualité, cohérence et intégrité — chacune avec un nombre, calculée sur la même extraction, et un tableau de bord qui résiste à la contestation.

Pierre Robentz Cassion

Leçon 1 sur 8

Évaluation de la qualité des données — Ce que qualité veut dire ici

Ce que couvre cette leçon

- La phrase qu'il faut cesser de dire
- Les cinq, et la mesure de chacune
- Calculez les cinq sur une seule extraction
- Pondérez les dimensions par ce qu'elles coûtent
- Ne réduisez pas à un seul nombre
- La suite

La phrase qu'il faut cesser de dire

- « La qualité des données est mauvaise.

Les cinq, et la mesure de chacune

Dimension	La question	La mesure
Complétude	Tout le monde a-t-il rapporté, et sur tout ?	Taux de rapportage ; taux de valeurs manquantes par champ
Actualité	Le rapport est-il arrivé à temps pour servir ?	Part transmise avant l'échéance ; retard médian en jours
Exactitude	Le chiffre remonté correspond-il à la source ?	Facteur de vérification — recomptage sur rapporté
Cohérence	Les nombres s'accordent-ils entre eux et avec le mois dernier ?	Taux d'échec des règles internes ; taux d'aberrations
Intégrité	La valeur vient-elle d'une mesure ou d'autre chose ?	Préférence de chiffres ; attraction ; séries anormalement stables

Les cinq, et la mesure de chacune

- L'exactitude est la seule qui exige le document source — Les quatre autres se calculent sur l'extraction que vous avez. . .
- L'intégrité n'est pas une accusation — Elle mesure si les nombres ressemblent à des mesures

Calculez les cinq sur une seule extraction — En Python (suite)

```
import pandas as pd

vax = pd.read_csv("vaccination-coverage-2024.v1.csv", parse_dates=["period"])
vax["reported"] = vax["report_submitted"] == True

scorecard = {}

# Completeness
scorecard["reporting_rate"] = vax["reported"].mean()
scorecard["field_completeness"] = 1 - vax.isna().mean().mean()

# Consistency: doses cannot exceed the target population
rule_fail = (vax["doses_administered"] > vax["target_population"]).mean()
scorecard["rule_failure_rate"] = rule_fail

# Integrity: share of reported values ending in zero
```

Calculez les cinq sur une seule extraction — En Python (suite)

```
reported = vax[vax["reported"]]
scorecard["round_number_share"] = (reported["doses_administered"] % 10 == 0).mean()

print(pd.Series(scorecard).round(3))
```

Calculez les cinq sur une seule extraction — En R

```
library(dplyr)
library(readr)

vax <- read_csv("vaccination-coverage-2024.v1.csv")

scorecard <- tibble::tibble(
  reporting_rate      = mean(vax$report_submitted),
  field_completeness = 1 - mean(is.na(as.matrix(vax))),
  rule_failure_rate   = mean(vax$doses_administered > vax$target_population),
  round_number_share = mean(
    vax$doses_administered[vax$report_submitted] %% 10 == 0
  )
)

round(scorecard, 3)
```

Calculez les cinq sur une seule extraction

Mesure	Valeur	À lire comme
Taux de rapportage	76,5 %	642 des 2 736 couples formation-mois ne portent aucun rapport
Complétude des champs	100 %	aucune case vide — ce qui n'est pas la même chose qu'aucune donnée manquante
Taux d'échec des règles	0 %	aucune formation ne rapporte plus de doses que sa population cible
Part de nombres ronds	10,4 %	ce que produit le hasard ; rien à voir

Pondérez les dimensions par ce qu'elles coûtent

- Pour un **taux de couverture**, la complétude domine. Un taux de rapportage de 76,5 % fait de tout total de district. . .
- Pour une **charge de cas servant à approvisionner**, l'exactitude domine. Commander des aliments thérapeutiques sur une. . .
- Pour un **indicateur d'alerte précoce**, l'actualité domine. Un rapport parfaitement exact arrivant six semaines après. . .
- Dites quelle dimension vous avez pondérée et pourquoi, dans le rapport — Un score composite qui cache sa pondération. . .

Ne réduisez pas à un seul nombre — En Python

```
scorecard_table = pd.DataFrame({
    "dimension": ["Completeness", "Timeliness", "Accuracy", "Consistency", "Integrity"],
    "measure": ["Reporting rate", "Submitted by deadline", "Verification factor",
               "Rule failure rate", "Round-number share"],
    "value": [0.765, None, None, 0.0, 0.104],
    "source": ["extract", "submission log", "facility visit", "extract", "extract"],
})
print(scorecard_table)
```

Ne réduisez pas à un seul nombre — En R

```
scorecard_table <- tibble::tribble(  
  ~dimension,    ~measure,                ~value, ~source,  
  "Completeness", "Reporting rate",          0.765, "extract",  
  "Timeliness",   "Submitted by deadline",    NA, "submission log",  
  "Accuracy",     "Verification factor",        NA, "facility visit",  
  "Consistency",  "Rule failure rate",          0.000, "extract",  
  "Integrity",    "Round-number share",        0.104, "extract"  
)
```

Ne réduisez pas à un seul nombre

- C'est la colonne source qui est utile — Elle dit immédiatement quels nombres vous pouvez produire aujourd'hui et. . .

- Deux des cinq dimensions ne sont presque jamais rapportées, et toutes deux se calculent sur des données que vous détenez déjà.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/data-quality-assessment/
dqa-01-five-dimensions](https://data-analysis.cassion.dev/fr/cours/data-quality-assessment/dqa-01-five-dimensions)