

Les fichiers Stata et SPSS gardent leurs étiquettes

haven, la classe labelled, et pourquoi le .dta d'un institut de sondage doit être lu sous R même si l'analyse se fait ailleurs — plus les deux façons de perdre un dictionnaire de codes sans s'en apercevoir.

Pierre Robentz Cassion

Leçon 3 sur 8

R et le tidyverse pour les données de programme — Des données qui gardent leur sens

Ce que couvre cette leçon

- Ce qu'un CSV jette
- La classe `labelled`
- Convertir délibérément
- Les deux façons de perdre un dictionnaire de codes
- Les valeurs manquantes définies par l'utilisateur
- SPSS et les autres formats
- Où cela rejoint le cours Python
- Ce qui vient ensuite

Ce qu'un CSV jette — En R

```
library(haven)
```

```
survey <- read_dta(here("data", "raw", "household-survey.dta"))
```

```
class(survey$consent)
#> [1] "haven_labelled" "vctrs_vctr" "double"

attr(survey$consent, "labels")
#>      Oui      Non Ne sait pas
#>      1      2      9

attr(survey$consent, "label")
#> [1] "Le chef de menage a consenti a l'entretien"
```

```
mean(survey$consent, na.rm = TRUE)  
#> [1] 1.7
```

- Une variable catégorielle devient un facteur

```
library(dplyr)

survey <- survey |>
  mutate(consent = as_factor(consent))

levels(survey$consent)
#> [1] "Oui" "Non" "Ne sait pas"
```

- Un véritable nombre perd ses étiquettes

```
survey <- survey |> mutate(hh_size = zap_labels(hh_size))
```

- Tout d'un coup, quand le fichier est uniformément catégoriel

```
survey <- read_dta(path) |> as_factor()
```

Les deux façons de perdre un dictionnaire de codes

- Lire avec les étiquettes déjà appliquées

Les deux façons de perdre un dictionnaire de codes — En R

```
survey <- read_dta(path, .name_repair = "unique")  
survey <- as_factor(survey)           # les valeurs ont disparu, restent les etiquettes
```

Les deux façons de perdre un dictionnaire de codes — En R

```
survey <- read_dta(path)

codes <- survey |> mutate(across(where(is.labelled), zap_labels))
labels <- survey |> as_factor()
```

Les deux façons de perdre un dictionnaire de codes

- Écrire le CSV et passer à autre chose

Les deux façons de perdre un dictionnaire de codes — En R

```
readr::write_csv(as_factor(survey), "survey.csv")
```

Les deux façons de perdre un dictionnaire de codes — En R

```
codebook <- tibble::tibble(  
  variable = names(survey),  
  label    = vapply(survey, function(x) attr(x, "label") %||% NA_character_, character(1))  
)  
  
readr::write_csv(codebook, "survey-codebook.csv")
```

```
survey$income  
#> <labelled<double>[3]>  
#> [1] 4500 NA(a) NA(b)
```

Les valeurs manquantes définies par l'utilisateur — En R

```
survey <- survey |> mutate(income = zap_missing(income))
```

```
read_sav(path)      # SPSS .sav  
read_por(path)      # SPSS portable  
read_dta(path)      # Stata  
read_xpt(path)      # SAS transport
```

```
write_dta(survey, "cleaned.dta")
```

- Le cours *Python pour les données de programme* dit, dans sa leçon sur les codes et les sentinelles, que c'est le seul endroit où la réponse est « utilisez l'autre langage ».

- Les étiquettes sont attachées et converties en facteurs.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/r-for-programme-data/
r-03-labelled-data](https://data-analysis.cassion.dev/fr/cours/r-for-programme-data/r-03-labelled-data)