Test de management et d’analyse des données avec R

Djerakei MISTALENGAR

2025-02-25

## Installation et chargement des packages

# Vérifier et installer les packages nécessaires  
  
packages <- c("haven", "utils", "dplyr", "tidyverse", "gtsummary", "survey", "knitr")  
  
for (pkg in packages) {  
 if (!require(pkg, character.only = TRUE)) install.packages(pkg, dependencies = TRUE)  
 library(pkg, character.only = TRUE)  
}

# Supprimer toutes les variables de l'environnement  
rm(list = ls())

## Chargement des données

# Chargement des fichiers  
  
mbl <- haven::read\_dta("../Données/food\_comp\_mother\_baseline.dta")  
  
mel <- haven::read\_dta("../Données/food\_comp\_mother\_endline.dta")  
  
  
str(mbl)

tibble [4,256 × 17] (S3: tbl\_df/tbl/data.frame) $ regionid : num [1:4256] 2 2 2 2 2 2 2 2 2 2 … ..- attr(*, “label”)= chr “Region ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ communeid : num [1:4256] 25 25 25 25 25 25 25 25 25 25 … ..- attr(*, “label”)= chr “Commune ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ villageid : num [1:4256] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 … ..- attr(*, “label”)= chr “Village ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ hhid : chr [1:4256] “4948484848535052” “4948484848535052” “4948484848535052” “4948484848535052” … ..- attr(*, “label”)= chr “Household ID” ..- attr(*, “format.stata”)= chr “%45s” $ round : dbl+lbl [1:4256] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1… ..@ label : chr “Survey round : Baseline, Endline” ..@ format.stata: chr “%10.0g” ..@ labels : Named num [1:2] 1 2 .. ..- attr(*, “names”)= chr [1:2] “Baseline” “Endline” $ s1\_q0 : dbl+lbl [1:4256] 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1… ..@ label : chr “eating occasion” ..@ format.stata: chr “%27.0g” ..@ labels : Named num [1:4] 1 2 3 4 .. ..- attr(*, “names”)= chr [1:4] “Breakfast” “Lunch” “Dinner” “Snacks” $ s1\_q1 : dbl+lbl [1:4256] 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1… ..@ label : chr “Meal consumed? Y/N” ..@ format.stata: chr “%9.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “No” “Yes” $ s1\_q2 : dbl+lbl [1:4256] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1… ..@ label : chr “Who consummed?” ..@ format.stata: chr “%16.0g” ..@ labels : Named num [1:3] 1 2 3 .. ..- attr(*, “names”)= chr [1:3] “mother” “child” “mother and child” $ V1 : num [1:4256] 680 NA 634 256 563 … ..- attr(*, “label”)= chr “Consommation en ernergie (kcal)” ..- attr(*, “format.stata”)= chr “%10.0g” $ protein\_g : num [1:4256] 23.31 NA 21.62 8.25 11.7 … ..- attr(*, “label”)= chr “Consommation en proteine (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ lipid\_tot\_g: num [1:4256] 5.3 NA 4.87 8.06 6.49 … ..- attr(*, “label”)= chr “Consommation en lipide (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ calcium\_mg : num [1:4256] 62.7 NA 57.1 22 116.3 … ..- attr(*, “label”)= chr “Consommation en calcium (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ iron\_mg : num [1:4256] 10.591 NA 9.897 0.912 2.716 … ..- attr(*, “label”)= chr “Consommation en fer (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ V9 : num [1:4256] 4.507 NA 4.19 0.456 3.382 … ..- attr(*, “label”)= chr “Consommation en zinc (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b6\_mg : num [1:4256] 0.3058 NA 0.2835 0.0456 0.2248 … ..- attr(*, “label”)= chr “Consommation en vitamine B6 (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b12\_mcg: num [1:4256] 0.00869 NA 0.0078 0 0.00823 … ..- attr(*, “label”)= chr “Consommation en vitamine B12 (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_c\_mg : num [1:4256] 0.0441 NA 0.0396 0 0.0002 … ..- attr(*, “label”)= chr “Consommation en vitamine C (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g”

str(mel)

tibble [4,256 × 17] (S3: tbl\_df/tbl/data.frame) $ regionid : num [1:4256] 2 2 2 2 2 2 2 2 2 2 … ..- attr(*, “label”)= chr “Region ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ communeid : num [1:4256] 25 25 25 25 25 25 25 25 25 25 … ..- attr(*, “label”)= chr “Commune ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ villageid : num [1:4256] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 … ..- attr(*, “label”)= chr “Village ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ hhid : chr [1:4256] “4948484848535052” “4948484848535052” “4948484848535052” “4948484848535052” … ..- attr(*, “label”)= chr “Household ID” ..- attr(*, “format.stata”)= chr “%45s” $ round : dbl+lbl [1:4256] 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2… ..@ label : chr “Survey round : Baseline, Endline” ..@ format.stata: chr “%10.0g” ..@ labels : Named num [1:2] 1 2 .. ..- attr(*, “names”)= chr [1:2] “Baseline” “Endline” $ s1\_q0 : dbl+lbl [1:4256] 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1… ..@ label : chr “eating occasion” ..@ format.stata: chr “%27.0g” ..@ labels : Named num [1:4] 1 2 3 4 .. ..- attr(*, “names”)= chr [1:4] “Breakfast” “Lunch” “Dinner” “Snacks” $ s1\_q1 : dbl+lbl [1:4256] 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1… ..@ label : chr “Meal consumed? Y/N” ..@ format.stata: chr “%9.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “No” “Yes” $ s1\_q2 : dbl+lbl [1:4256] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1… ..@ label : chr “Who consummed?” ..@ format.stata: chr “%16.0g” ..@ labels : Named num [1:3] 1 2 3 .. ..- attr(*, “names”)= chr [1:3] “mother” “child” “mother and child” $ energ\_kcal : num [1:4256] 2217 1021 3038 716 618 … ..- attr(*, “label”)= chr “Consommation en ernergie (kcal)” ..- attr(*, “format.stata”)= chr “%10.0g” $ protein\_g : num [1:4256] 48.7 34.8 92 19.1 16.5 … ..- attr(*, “label”)= chr “Consommation en proteine (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ lipid\_tot\_g: num [1:4256] 89.6 5.94 183.6 4.73 1.63 … ..- attr(*, “label”)= chr “Consommation en lipide (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ calcium\_mg : num [1:4256] 818.8 72.1 1781.3 131.6 10.8 … ..- attr(*, “label”)= chr “Consommation en calcium (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ iron\_mg : num [1:4256] 22.659 16.558 11.345 8.877 0.105 … ..- attr(*, “label”)= chr “Consommation en fer (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ zinc\_mg : num [1:4256] 9.5691 6.4618 2.2278 3.5867 0.0334 … ..- attr(*, “label”)= chr “Consommation en zinc (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b6\_mg : num [1:4256] 0.8748 0.4327 0.314 0.5277 0.0059 … ..- attr(*, “label”)= chr “Consommation en vitamine B6 (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b12\_mcg: num [1:4256] 0.00948 0.02232 0 0.01281 0.01016 … ..- attr(*, “label”)= chr “Consommation en vitamine B12 (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_c\_mg : num [1:4256] 11.547 0.183 26.738 0.133 0 … ..- attr(*, “label”)= chr “Consommation en vitamine C (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g”

cbl <- haven::read\_dta("../Données/food\_comp\_child\_baseline.dta")  
  
cel <- haven::read\_dta("../Données/food\_comp\_child\_endline.dta")  
  
  
str(cbl)

tibble [4,256 × 17] (S3: tbl\_df/tbl/data.frame) $ regionid : num [1:4256] 2 2 2 2 2 2 2 2 2 2 … ..- attr(*, “label”)= chr “Region ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ communeid : num [1:4256] 25 25 25 25 25 25 25 25 25 25 … ..- attr(*, “label”)= chr “Commune ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ villageid : num [1:4256] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 … ..- attr(*, “label”)= chr “Village ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ hhid : chr [1:4256] “4948484848535052” “4948484848535052” “4948484848535052” “4948484848535052” … ..- attr(*, “label”)= chr “Household ID” ..- attr(*, “format.stata”)= chr “%45s” $ round : dbl+lbl [1:4256] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1… ..@ label : chr “Survey round : Baseline, Endline” ..@ format.stata: chr “%10.0g” ..@ labels : Named num [1:2] 1 2 .. ..- attr(*, “names”)= chr [1:2] “Baseline” “Endline” $ s1\_q0 : dbl+lbl [1:4256] 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1… ..@ label : chr “eating occasion” ..@ format.stata: chr “%27.0g” ..@ labels : Named num [1:4] 1 2 3 4 .. ..- attr(*, “names”)= chr [1:4] “Breakfast” “Lunch” “Dinner” “Snacks” $ s1\_q1 : dbl+lbl [1:4256] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1… ..@ label : chr “Meal consumed? Y/N” ..@ format.stata: chr “%9.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “No” “Yes” $ s1\_q2 : dbl+lbl [1:4256] 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2… ..@ label : chr “Who consummed?” ..@ format.stata: chr “%16.0g” ..@ labels : Named num [1:3] 1 2 3 .. ..- attr(*, “names”)= chr [1:3] “mother” “child” “mother and child” $ energ\_kcal : num [1:4256] 355 224 334 494 235 … ..- attr(*, “label”)= chr “Consommation en ernergie (kcal)” ..- attr(*, “format.stata”)= chr “%10.0g” $ protein\_g : num [1:4256] 12.74 8.01 12.72 16.65 4.45 … ..- attr(*, “label”)= chr “Consommation en proteine (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ lipid\_tot\_g: num [1:4256] 3.14 1.96 3.42 8.84 4.76 … ..- attr(*, “label”)= chr “Consommation en lipide (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ calcium\_mg : num [1:4256] 40.1 24.8 47 60.1 93.3 … ..- attr(*, “label”)= chr “Consommation en calcium (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ iron\_mg : num [1:4256] 5.39 3.42 4.9 6.83 1.54 … ..- attr(*, “label”)= chr “Consommation en fer (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ zinc\_mg : num [1:4256] 2.41 1.52 2.35 3.13 1.63 … ..- attr(*, “label”)= chr “Consommation en zinc (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b6\_mg : num [1:4256] 0.167 0.105 0.168 0.207 0.116 … ..- attr(*, “label”)= chr “Consommation en vitamine B6 (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b12\_mcg: num [1:4256] 0.00616 0.00377 0.00785 0.00959 0.00688 … ..- attr(*, “label”)= chr “Consommation en vitamine B12 (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_c\_mg : num [1:4256] 0.031282 0.019173 0.03986 0.048689 0.000167 … ..- attr(*, “label”)= chr “Consommation en vitamine C (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g”

str(cel)

tibble [4,256 × 17] (S3: tbl\_df/tbl/data.frame) $ regionid : num [1:4256] 2 2 2 2 2 2 2 2 2 2 … ..- attr(*, “label”)= chr “Region ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ communeid : num [1:4256] 25 25 25 25 25 25 25 25 25 25 … ..- attr(*, “label”)= chr “Commune ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ villageid : num [1:4256] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 … ..- attr(*, “label”)= chr “Village ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ hhid : chr [1:4256] “4948484848535052” “4948484848535052” “4948484848535052” “4948484848535052” … ..- attr(*, “label”)= chr “Household ID” ..- attr(*, “format.stata”)= chr “%45s” $ round : dbl+lbl [1:4256] 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2… ..@ label : chr “Survey round : Baseline, Endline” ..@ format.stata: chr “%10.0g” ..@ labels : Named num [1:2] 1 2 .. ..- attr(*, “names”)= chr [1:2] “Baseline” “Endline” $ s1\_q0 : dbl+lbl [1:4256] 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1… ..@ label : chr “eating occasion” ..@ format.stata: chr “%27.0g” ..@ labels : Named num [1:4] 1 2 3 4 .. ..- attr(*, “names”)= chr [1:4] “Breakfast” “Lunch” “Dinner” “Snacks” $ s1\_q1 : dbl+lbl [1:4256] 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1… ..@ label : chr “Meal consumed? Y/N” ..@ format.stata: chr “%9.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “No” “Yes” $ s1\_q2 : dbl+lbl [1:4256] 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2… ..@ label : chr “Who consummed?” ..@ format.stata: chr “%16.0g” ..@ labels : Named num [1:3] 1 2 3 .. ..- attr(*, “names”)= chr [1:3] “mother” “child” “mother and child” $ energ\_kcal : num [1:4256] 1193 535 1701 531 188 … ..- attr(*, “label”)= chr “Consommation en ernergie (kcal)” ..- attr(*, “format.stata”)= chr “%10.0g” $ protein\_g : num [1:4256] 24.94 18.38 58.84 13.89 5.04 … ..- attr(*, “label”)= chr “Consommation en proteine (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ lipid\_tot\_g: num [1:4256] 51.915 3.205 90.268 3.917 0.496 … ..- attr(*, “label”)= chr “Consommation en lipide (g)” ..- attr(*, “format.stata”)= chr “%10.0g” $ calcium\_mg : num [1:4256] 479.73 41.32 1122.63 74.44 3.29 … ..- attr(*, “label”)= chr “Consommation en calcium (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ iron\_mg : num [1:4256] 11.5857 8.709 7.2277 6.4328 0.0321 … ..- attr(*, “label”)= chr “Consommation en fer (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ zinc\_mg : num [1:4256] 4.9181 3.3562 1.4184 2.7147 0.0102 … ..- attr(*, “label”)= chr “Consommation en zinc (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b6\_mg : num [1:4256] 0.4678 0.2251 0.2028 0.3523 0.0018 … ..- attr(*, “label”)= chr “Consommation en vitamine B6 (mg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_b12\_mcg: num [1:4256] 0.00396 0.01364 0 0.00736 0.00309 … ..- attr(*, “label”)= chr “Consommation en vitamine B12 (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g” $ vit\_c\_mg : num [1:4256] 6.8299 0.1201 16.904 0.0764 0 … ..- attr(*, “label”)= chr “Consommation en vitamine C (mcg)” ..- attr(*, “format.stata”)= chr “%10.0g”

men <- haven::read\_dta("../Données/base\_menage.dta")  
  
str(men)

tibble [1,065 × 21] (S3: tbl\_df/tbl/data.frame) $ regionid : num [1:1065] 2 2 2 2 2 2 2 2 2 2 … ..- attr(*, “label”)= chr “Region ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ communeid : num [1:1065] 25 25 25 25 25 25 25 25 25 25 … ..- attr(*, “label”)= chr “Commune ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ villageid : num [1:1065] 1000 1000 1000 1000 1000 1000 1000 1000 1000 1000 … ..- attr(*, “label”)= chr “Village ID” ..- attr(*, “format.stata”)= chr “%8.0g” $ hhid : chr [1:1065] “4948484848535052” “4948484848535053” “4948484848535055” “4948484848535056” … ..- attr(*, “label”)= chr “Household ID” ..- attr(*, “format.stata”)= chr “%45s” $ hhsize : num [1:1065] 4 8 11 9 16 6 31 8 23 5 … ..- attr(*, “label”)= chr “Household size” ..- attr(*, “format.stata”)= chr “%10.0g” $ poly : dbl+lbl [1:1065] 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1… ..@ label : chr “Polygamous household?” ..@ format.stata: chr “%8.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Non” “Oui” $ hh\_primary : dbl+lbl [1:1065] 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0… ..@ label : chr “Household head completed primary education” ..@ format.stata: chr “%9.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Non” “Oui” $ s1\_q2 : dbl+lbl [1:1065] 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1… ..@ label : chr “Male head of household” ..@ format.stata: chr “%8.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Female” “Male” $ s1\_q4a : num [1:1065] 37 52 67 38 75 41 56 52 63 26 … ..- attr(*, “label”)= chr “Age head of household” ..- attr(*, “format.stata”)= chr “%8.0g” $ s2\_q1 : dbl+lbl [1:1065] 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0… ..@ label : chr “Is head of household literate in local language” ..@ format.stata: chr “%8.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Non” “Oui” $ s2\_q2 : dbl+lbl [1:1065] 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0… ..@ label : chr “Is head of household literate in French?” ..@ format.stata: chr “%8.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Non” “Oui” $ s2\_q4 : dbl+lbl [1:1065] 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1… ..@ label : chr “Has head of household been to school (formal or informal)?” ..@ format.stata: chr “%8.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Non” “Oui” $ s29\_q1 : dbl+lbl [1:1065] 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0… ..@ label : chr “Est-ce qu’un membre de votre ménage a pris un prêt ou fait un emprunt en argent” ..@ format.stata: chr “%8.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Non” “Oui” $ demgrp1 : num [1:1065] 1 0 0 0 1 0 5 1 1 1 … ..- attr(*, “label”)= chr “Number of children 0-36 months” ..- attr(*, “format.stata”)= chr “%9.0g” $ demgrp2 : num [1:1065] 1 2 2 1 1 1 5 0 2 1 … ..- attr(*, “label”)= chr “Number of children 36-72 months” ..- attr(*, “format.stata”)= chr “%9.0g” $ demgrp3 : num [1:1065] 0 3 3 3 7 2 11 4 9 1 … ..- attr(*, “label”)= chr “Number of adults 6-14 years” ..- attr(*, “format.stata”)= chr “%9.0g” $ demgrp4 : num [1:1065] 1 2 5 5 6 3 10 3 11 2 … ..- attr(*, “label”)= chr “Number of adults 14-65 years” ..- attr(*, “format.stata”)= chr “%9.0g” $ demgrp5 : num [1:1065] 0 1 1 0 1 0 0 0 0 0 … ..- attr(*, “label”)= chr “Number of elders 65+ years” ..- attr(*, “format.stata”)= chr “%9.0g” $ dependencyratio: num [1:1065] 2 3 1.2 0.8 1.67 … ..- attr(*, “label”)= chr “Dependency ratio” ..- attr(*, “format.stata”)= chr “%9.0g” $ hfias\_score : num [1:1065] 18 21 0 4 1 3 12 0 14 0 … ..- attr(*, “label”)= chr “HFIAS Score (0-27)” ..- attr(*, “format.stata”)= chr “%9.0g” $ T1 : dbl+lbl [1:1065] 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1… ..@ label : chr “Treatment: 1st level comparison” ..@ format.stata: chr “%12.0g” ..@ labels : Named num [1:2] 0 1 .. ..- attr(*, “names”)= chr [1:2] “Controle” “Intervention” - attr(\*, “label”)= chr “One row per household”

## Partie 1 : Gestion et nettoyage des bases de données

#### 1. Vérifiez la présence de doublons dans les bases de données Baseline, Endline et ménage. Supprimez les doublons si nécessaire.

colnames(mbl)

[1] “regionid” “communeid” “villageid” “hhid” “round”  
[6] “s1\_q0” “s1\_q1” “s1\_q2” “V1” “protein\_g”  
[11] “lipid\_tot\_g” “calcium\_mg” “iron\_mg” “V9” “vit\_b6\_mg”  
[16] “vit\_b12\_mcg” “vit\_c\_mg”

View(mbl)

# Fonction pour vérifier et supprimer les doublons  
clean\_data <- function(df, key\_vars) {  
 df <- df %>%  
 mutate(dupli = duplicated(df[key\_vars])) # Identifier les doublons  
 print(sum(df$dupli)) # Nombre de doublons  
   
 df <- df %>%  
 distinct(across(all\_of(key\_vars)), .keep\_all = TRUE) # Supprimer les doublons  
   
 return(df)  
}

mbl<-clean\_data(mbl, "hhid")

[1] 3192

print(mbl)

mel<-clean\_data(mel, "hhid")

[1] 3192

cbl<-clean\_data(cbl, "hhid")

[1] 3192

cbl

cel<-clean\_data(cel, "hhid")

[1] 3192

# Vérification des noms de variables dans chaque base de données  
names(mbl) # Baseline des mères

[1] “regionid” “communeid” “villageid” “hhid” “round”  
[6] “s1\_q0” “s1\_q1” “s1\_q2” “V1” “protein\_g”  
[11] “lipid\_tot\_g” “calcium\_mg” “iron\_mg” “V9” “vit\_b6\_mg”  
[16] “vit\_b12\_mcg” “vit\_c\_mg” “dupli”

names(mel) # Endline des mères

[1] “regionid” “communeid” “villageid” “hhid” “round”  
[6] “s1\_q0” “s1\_q1” “s1\_q2” “energ\_kcal” “protein\_g”  
[11] “lipid\_tot\_g” “calcium\_mg” “iron\_mg” “zinc\_mg” “vit\_b6\_mg”  
[16] “vit\_b12\_mcg” “vit\_c\_mg” “dupli”

names(cbl) # Baseline des enfants

[1] “regionid” “communeid” “villageid” “hhid” “round”  
[6] “s1\_q0” “s1\_q1” “s1\_q2” “energ\_kcal” “protein\_g”  
[11] “lipid\_tot\_g” “calcium\_mg” “iron\_mg” “zinc\_mg” “vit\_b6\_mg”  
[16] “vit\_b12\_mcg” “vit\_c\_mg” “dupli”

names(cel) # Endline des enfants

[1] “regionid” “communeid” “villageid” “hhid” “round”  
[6] “s1\_q0” “s1\_q1” “s1\_q2” “energ\_kcal” “protein\_g”  
[11] “lipid\_tot\_g” “calcium\_mg” “iron\_mg” “zinc\_mg” “vit\_b6\_mg”  
[16] “vit\_b12\_mcg” “vit\_c\_mg” “dupli”

names(men) # Base des ménages

[1] “regionid” “communeid” “villageid” “hhid”  
[5] “hhsize” “poly” “hh\_primary” “s1\_q2”  
[9] “s1\_q4a” “s2\_q1” “s2\_q2” “s2\_q4”  
[13] “s29\_q1” “demgrp1” “demgrp2” “demgrp3”  
[17] “demgrp4” “demgrp5” “dependencyratio” “hfias\_score”  
[21] “T1”

#### 2. Assurez-vous que les noms des variables sont cohérents entre les bases de données Baseline et Endline

# Comparer les noms de variables pour identifier les différences  
  
# Comparer les noms entre Baseline et Endline pour les mères  
setdiff(names(mbl), names(mel))

[1] “V1” “V9”

setdiff(names(mel), names(mbl))

[1] “energ\_kcal” “zinc\_mg”

# Comparer les noms entre Baseline et Endline pour les enfants  
setdiff(names(cbl), names(cel))

character(0)

setdiff(names(cel), names(cbl))

character(0)

# Renommer la colonne 'V1' en 'energ\_kcal'  
  
colnames(mel)[colnames(mel) == "V1"] <- "energ\_kcal"  
  
colnames(mbl)[colnames(mbl) == "V1"] <- "energ\_kcal"  
  
# Vérifier les noms des variables après renaming  
names(mbl)

[1] “regionid” “communeid” “villageid” “hhid” “round”  
[6] “s1\_q0” “s1\_q1” “s1\_q2” “energ\_kcal” “protein\_g”  
[11] “lipid\_tot\_g” “calcium\_mg” “iron\_mg” “V9” “vit\_b6\_mg”  
[16] “vit\_b12\_mcg” “vit\_c\_mg” “dupli”

names(mel)

[1] “regionid” “communeid” “villageid” “hhid” “round”  
[6] “s1\_q0” “s1\_q1” “s1\_q2” “energ\_kcal” “protein\_g”  
[11] “lipid\_tot\_g” “calcium\_mg” “iron\_mg” “zinc\_mg” “vit\_b6\_mg”  
[16] “vit\_b12\_mcg” “vit\_c\_mg” “dupli”

#### Veuillez vérifier soigneusement les données et corriger les données manquantes de certaines variables si possibles.

# Vérifier les données manquantes dans chaque base  
sum(is.na(mbl)) # Pour Baseline des mères

[1] 615

sum(is.na(mel)) # Pour Endline des mères

[1] 702

sum(is.na(cbl)) # Pour Baseline des enfants

[1] 306

sum(is.na(cel)) # Pour Endline des enfants

[1] 486

sum(is.na(men)) # Pour la base des ménages

[1] 1

# Pour la base `mbl` (Baseline des mères), par exemple  
colSums(is.na(mbl)) # Compter les NA pour chaque variable

regionid communeid villageid hhid round s1\_q0 1 2 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 68 68 68 68 iron\_mg V9 vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 68 68 68 68 68 0

colSums(is.na(mel)) # Idem pour la base `mel` (Endline des mères)

regionid communeid villageid hhid round s1\_q0 0 0 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 78 78 78 78 iron\_mg zinc\_mg vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 78 78 78 78 78 0

# Répéter pour les autres bases  
colSums(is.na(cbl))

regionid communeid villageid hhid round s1\_q0 0 0 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 34 34 34 34 iron\_mg zinc\_mg vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 34 34 34 34 34 0

colSums(is.na(cel))

regionid communeid villageid hhid round s1\_q0 0 0 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 54 54 54 54 iron\_mg zinc\_mg vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 54 54 54 54 54 0

colSums(is.na(men))

regionid communeid villageid hhid   
 0 1 0 0   
 hhsize poly hh\_primary s1\_q2   
 0 0 0 0   
 s1\_q4a s2\_q1 s2\_q2 s2\_q4   
 0 0 0 0   
 s29\_q1 demgrp1 demgrp2 demgrp3   
 0 0 0 0   
 demgrp4 demgrp5 dependencyratio hfias\_score   
 0 0 0 0   
 T1   
 0

# Imputation par la médiane pour les nutriments  
mbl$protein\_g[is.na(mbl$protein\_g)] <- median(mbl$protein\_g, na.rm = TRUE)  
mel$protein\_g[is.na(mel$protein\_g)] <- median(mel$protein\_g, na.rm = TRUE)  
  
mbl$lipid\_tot\_g[is.na(mbl$lipid\_tot\_g)] <- median(mbl$lipid\_tot\_g, na.rm = TRUE)  
mel$lipid\_tot\_g[is.na(mel$lipid\_tot\_g)] <- median(mel$lipid\_tot\_g, na.rm = TRUE)  
  
# Imputation par 0 pour les variables de type 'absence de consommation' comme `s1\_q0`  
mbl$s1\_q0[is.na(mbl$s1\_q0)] <- 0  
mel$s1\_q0[is.na(mel$s1\_q0)] <- 0  
  
# Vérification après traitement  
colSums(is.na(mbl)) # Vérifier les NA après traitement

regionid communeid villageid hhid round s1\_q0 1 2 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 68 0 0 68 iron\_mg V9 vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 68 68 68 68 68 0

colSums(is.na(mel))

regionid communeid villageid hhid round s1\_q0 0 0 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 78 0 0 78 iron\_mg zinc\_mg vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 78 78 78 78 78 0

# Imputation dans les données des enfants  
cbl$protein\_g[is.na(cbl$protein\_g)] <- median(cbl$protein\_g, na.rm = TRUE)  
cel$protein\_g[is.na(cel$protein\_g)] <- median(cel$protein\_g, na.rm = TRUE)  
  
# Imputation des valeurs manquantes pour les nutriments spécifiques  
cbl$calcium\_mg[is.na(cbl$calcium\_mg)] <- median(cbl$calcium\_mg, na.rm = TRUE)  
cbl$iron\_mg[is.na(cbl$iron\_mg)] <- median(cbl$iron\_mg, na.rm = TRUE)  
cbl$zinc\_mg[is.na(cbl$zinc\_mg)] <- median(cbl$zinc\_mg, na.rm = TRUE)  
cbl$vit\_b6\_mg[is.na(cbl$vit\_b6\_mg)] <- median(cbl$vit\_b6\_mg, na.rm = TRUE)  
cbl$vit\_b12\_mcg[is.na(cbl$vit\_b12\_mcg)] <- median(cbl$vit\_b12\_mcg, na.rm = TRUE)  
cbl$vit\_c\_mg[is.na(cbl$vit\_c\_mg)] <- median(cbl$vit\_c\_mg, na.rm = TRUE)  
  
  
# Imputation par la médiane pour les variables avec des NA  
cbl$energ\_kcal[is.na(cbl$energ\_kcal)] <- median(cbl$energ\_kcal, na.rm = TRUE)  
cel$calcium\_mg[is.na(cel$calcium\_mg)] <- median(cel$calcium\_mg, na.rm = TRUE)  
# Imputation par la médiane pour les nutriments manquants  
cel$iron\_mg[is.na(cel$iron\_mg)] <- median(cel$iron\_mg, na.rm = TRUE)  
cel$zinc\_mg[is.na(cel$zinc\_mg)] <- median(cel$zinc\_mg, na.rm = TRUE)  
cel$vit\_b6\_mg[is.na(cel$vit\_b6\_mg)] <- median(cel$vit\_b6\_mg, na.rm = TRUE)  
cel$vit\_b12\_mcg[is.na(cel$vit\_b12\_mcg)] <- median(cel$vit\_b12\_mcg, na.rm = TRUE)  
cel$vit\_c\_mg[is.na(cel$vit\_c\_mg)] <- median(cel$vit\_c\_mg, na.rm = TRUE)  
  
  
cbl$lipid\_tot\_g[is.na(cbl$lipid\_tot\_g)] <- median(cbl$lipid\_tot\_g, na.rm = TRUE)  
cel$lipid\_tot\_g[is.na(cel$lipid\_tot\_g)] <- median(cel$lipid\_tot\_g, na.rm = TRUE)  
  
# Et répéter pour d'autres variables comme calcium, zinc, etc.  
  
  
# Imputation dans la base des ménages  
men$hhsize[is.na(men$hhsize)] <- median(men$hhsize, na.rm = TRUE)  
  
  
# Vérification après traitement  
colSums(is.na(cbl)) # Vérifier les NA après traitement

regionid communeid villageid hhid round s1\_q0 0 0 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 0 0 0 0 iron\_mg zinc\_mg vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 0 0 0 0 0 0

colSums(is.na(cel))

regionid communeid villageid hhid round s1\_q0 0 0 0 0 0 0 s1\_q1 s1\_q2 energ\_kcal protein\_g lipid\_tot\_g calcium\_mg 0 0 54 0 0 0 iron\_mg zinc\_mg vit\_b6\_mg vit\_b12\_mcg vit\_c\_mg dupli 0 0 0 0 0 0

colSums(is.na(men))

regionid communeid villageid hhid   
 0 1 0 0   
 hhsize poly hh\_primary s1\_q2   
 0 0 0 0   
 s1\_q4a s2\_q1 s2\_q2 s2\_q4   
 0 0 0 0   
 s29\_q1 demgrp1 demgrp2 demgrp3   
 0 0 0 0   
 demgrp4 demgrp5 dependencyratio hfias\_score   
 0 0 0 0   
 T1   
 0

#### 4. La consommation d’énergie moyenne à chaque repas pour l’ensemble des mères lors de l’enquête de base

On va calculer la moyenne de la consommation d’énergie pour l’ensemble des repas dans l’enquête de base.

# Calculer la consommation moyenne d'énergie en ignorant les valeurs manquantes  
  
mean\_energ <- mean(mbl$energ\_kcal, na.rm = TRUE)  
  
# Afficher la moyenne  
mean\_energ

[1] 636.7127

# Sauvegarder les changements dans les fichiers modifiés  
  
# Baseline des mères  
haven::write\_dta(mbl, "../Données/mother\_baseline\_v1.dta")  
  
# Endline des mères  
haven::write\_dta(mel, "../Données/mother\_endline\_v1.dta")   
  
 # Baseline des enfants  
haven::write\_dta(cbl, "../Données/child\_baseline\_v1.dta")  
  
# Endline des enfants  
haven::write\_dta(cel, "../Données/child\_endline\_v1.dta")   
  
 # Base des ménages  
haven::write\_dta(men, "../Données/base\_menage\_final.dta")

# Partie 2 : Empilement et Fusion des données

## 1) Baseline

### i) Empilez les bases de données

# Charger les bases de données à nouveau  
  
mbl <- haven::read\_dta("../Données/mother\_baseline\_v1.dta")  
cbl <- haven::read\_dta("../Données/child\_baseline\_v1.dta")

# Effectuer un left join (on veut garder toutes les lignes de la mère)  
merged\_data <- right\_join(cbl, mbl, by = "hhid")  
  
  
# Voir  
View(merged\_data)

### ii) Renommage de toutes les variables de consommation energ\_kcal jusqu’à vit\_c\_mcg en ajoutant le suffixe \_b pour faire référence à l’enquête Baseline.

merged\_data <- merged\_data %>%  
 rename\_with(~ gsub("\\.x$", "\_b", .), contains(c("energ\_kcal", "protein\_g", "lipid\_tot\_g", "calcium\_mg", "iron\_mg", "vit\_b6\_mg", "vit\_b12\_mcg", "vit\_c\_mg")))  
  
# Renommer les colonnes de consommation d'enfants avec le suffixe "\_c" pour l'enquête Baseline  
  
  
merged\_data <- merged\_data %>%  
 rename\_with(~ gsub("\\.y$", "\_c", .), contains(c("energ\_kcal", "protein\_g", "lipid\_tot\_g", "calcium\_mg", "iron\_mg", "vit\_b6\_mg", "vit\_b12\_mcg", "vit\_c\_mg")))  
  
# Vérifier les noms de colonnes après renommage  
names(merged\_data)

[1] “regionid.x” “communeid.x” “villageid.x” “hhid”  
[5] “round.x” “s1\_q0.x” “s1\_q1.x” “s1\_q2.x”  
[9] “energ\_kcal\_b” “protein\_g\_b” “lipid\_tot\_g\_b” “calcium\_mg\_b” [13] “iron\_mg\_b” “zinc\_mg” “vit\_b6\_mg\_b” “vit\_b12\_mcg\_b” [17] “vit\_c\_mg\_b” “dupli.x” “regionid.y” “communeid.y”  
[21] “villageid.y” “round.y” “s1\_q0.y” “s1\_q1.y”  
[25] “s1\_q2.y” “energ\_kcal\_c” “protein\_g\_c” “lipid\_tot\_g\_c” [29] “calcium\_mg\_c” “iron\_mg\_c” “V9” “vit\_b6\_mg\_c”  
[33] “vit\_b12\_mcg\_c” “vit\_c\_mg\_c” “dupli.y”

### iii) Création d’une base de données qui résume les consommations journalières totales par individu (somme des 4 repas) pour l’énergie et tous les nutriments en utilisant la commande merge.

# Créer un résumé des consommations journalières totales  
summary\_data <- merged\_data %>%  
 mutate(  
 total\_energ\_kcal = energ\_kcal\_b + energ\_kcal\_c,  
 total\_protein\_g = protein\_g\_b + protein\_g\_c,  
 total\_lipid\_g = lipid\_tot\_g\_b + lipid\_tot\_g\_c,  
 total\_calcium\_mg = calcium\_mg\_b + calcium\_mg\_c,  
 total\_iron\_mg = iron\_mg\_b + iron\_mg\_c,  
 total\_zinc\_mg = zinc\_mg,   
 total\_vit\_b6\_mg = vit\_b6\_mg\_b + vit\_b6\_mg\_c,  
 total\_vit\_b12\_mcg = vit\_b12\_mcg\_b + vit\_b12\_mcg\_c,  
 total\_vit\_c\_mg = vit\_c\_mg\_b + vit\_c\_mg\_c  
 ) %>%  
 select(hhid, total\_energ\_kcal, total\_protein\_g, total\_lipid\_g, total\_calcium\_mg,   
 total\_iron\_mg, total\_zinc\_mg, total\_vit\_b6\_mg, total\_vit\_b12\_mcg, total\_vit\_c\_mg)  
  
# Vérifier le résultat  
head(summary\_data)  
  
# Sauvegarder la base de données résumée  
haven::write\_dta(summary\_data, "../Données/summary\_daily\_consumption.dta")

### iv) Sauvegarde la base de données finale

# Sélectionner les colonnes nécessaires  
baseline\_final <- merged\_data %>%  
 select(hhid, s1\_q2.x, energ\_kcal\_b, protein\_g\_b, lipid\_tot\_g\_b, calcium\_mg\_b, iron\_mg\_b, zinc\_mg) %>%  
 rename(  
 s1\_q2 = s1\_q2.x,  
 energ\_kcal = energ\_kcal\_b,  
 protein\_g = protein\_g\_b,  
 lipid\_tot\_g = lipid\_tot\_g\_b,  
 calcium\_mg = calcium\_mg\_b,  
 iron\_mg = iron\_mg\_b  
 )  
  
# Vérifier le résultat  
head(baseline\_final)  
  
# Sauvegarder la base de données finale  
haven::write\_dta(baseline\_final, "../Données/baseline\_final.dta")

# Charger les bases de données Endline  
mel <- haven::read\_dta("../Données/mother\_endline\_v1.dta")  
cel <- haven::read\_dta("../Données/child\_endline\_v1.dta")  
  
# Renommer les variables de consommation pour l'enquête Endline en ajoutant le suffixe "\_e"  
mel <- mel %>%  
 rename(  
 energ\_kcal\_e = energ\_kcal,  
 protein\_g\_e = protein\_g,  
 lipid\_tot\_g\_e = lipid\_tot\_g,  
 calcium\_mg\_e = calcium\_mg,  
 iron\_mg\_e = iron\_mg,  
 zinc\_mg\_e = zinc\_mg,  
 vit\_b6\_mg\_e = vit\_b6\_mg,  
 vit\_b12\_mcg\_e = vit\_b12\_mcg,  
 vit\_c\_mg\_e = vit\_c\_mg  
 )  
  
cel <- cel %>%  
 rename(  
 energ\_kcal\_e = energ\_kcal,  
 protein\_g\_e = protein\_g,  
 lipid\_tot\_g\_e = lipid\_tot\_g,  
 calcium\_mg\_e = calcium\_mg,  
 iron\_mg\_e = iron\_mg,  
 zinc\_mg\_e = zinc\_mg,  
 vit\_b6\_mg\_e = vit\_b6\_mg,  
 vit\_b12\_mcg\_e = vit\_b12\_mcg,  
 vit\_c\_mg\_e = vit\_c\_mg  
 )  
  
# Empiler les données des mères et des enfants pour chaque ménage  
endline\_merged <- bind\_rows(mel, cel)  
  
# Vérifier la base fusionnée  
head(endline\_merged)  
  
# Sauvegarder la base de données empilée sous le nom "endline\_merged\_mother\_child.dta"  
haven::write\_dta(endline\_merged, "../Données/endline\_merged\_mother\_child.dta")

## Endline

# Charger les bases de données des mères et des enfants  
mel <- haven::read\_dta("../Données/mother\_endline\_v1.dta")  
cel <- haven::read\_dta("../Données/child\_endline\_v1.dta")  
  
# Fusionner les bases de données des mères et des enfants sur "hhid" (identifiant du ménage)  
merged\_endline <- right\_join( cel,mel, by = "hhid")

colnames(merged\_endline)

[1] “regionid.x” “communeid.x” “villageid.x” “hhid”  
[5] “round.x” “s1\_q0.x” “s1\_q1.x” “s1\_q2.x”  
[9] “energ\_kcal.x” “protein\_g.x” “lipid\_tot\_g.x” “calcium\_mg.x” [13] “iron\_mg.x” “zinc\_mg.x” “vit\_b6\_mg.x” “vit\_b12\_mcg.x” [17] “vit\_c\_mg.x” “dupli.x” “regionid.y” “communeid.y”  
[21] “villageid.y” “round.y” “s1\_q0.y” “s1\_q1.y”  
[25] “s1\_q2.y” “energ\_kcal.y” “protein\_g.y” “lipid\_tot\_g.y” [29] “calcium\_mg.y” “iron\_mg.y” “zinc\_mg.y” “vit\_b6\_mg.y”  
[33] “vit\_b12\_mcg.y” “vit\_c\_mg.y” “dupli.y”

### 3) Fusionnez les données baseline\_final.dta et endline\_final.dta

# Charger les deux bases de données  
baseline\_data <- haven::read\_dta("../Données/baseline\_final.dta")  
endline\_data <- haven::read\_dta("../Données/endline\_final.dta")  
  
# Fusionner les données Baseline et Endline par hhid  
merged\_data <- left\_join(baseline\_data, endline\_data, by = "hhid", suffix = c("\_b", "\_e"))  
  
# Vérifier le résultat de la fusion  
head(merged\_data)  
  
# Sauvegarder la base de données fusionnée  
haven::write\_dta(merged\_data, "../Données/merged\_baseline\_endline.dta")

# Charger la base  
base\_menage <- read\_dta("../Données/base\_menage.dta")

#Taille du menage  
base\_household\_size <- base\_menage %>%   
 select(hhid, hhsize) %>%   
 distinct()

# Education  
base\_education\_level <- base\_menage %>%   
 select(hhid, hh\_primary) %>%   
 distinct()

# Ratio  
  
base\_dependence\_ratio <- base\_menage %>%   
 select(hhid, dependencyratio) %>%   
 distinct()

# sécurité alimentaire)  
  
base\_HFIAS\_score <- base\_menage %>%   
 select(hhid, hfias\_score) %>%   
 distinct()