RÉPUBLIQUE DU SÉNÉGAL

Un Peuple - Un But - Une Foi

Ministère de l'Économie, du Plan et de la Coopération

****

Agence nationale de la Statistique et de la Démographie (ANSD)

École nationale de la Statistique et de l'Analyse économique Pierre Ndiaye (ENSAE)

PROJET STATISTIQUES SOUS R

TP9 : Fusion des bases welfare des EHCVM 2018 ET 2021 du Sénégal

Rédigé par :

Jean Luc BATABATI

Élève ingénieur statisticien économiste

Sous la supervision de :

M. Aboubacar HEMA

Analyste de recherche chez IFPRI

Année scolaire : 2024/2025

Contents

L'objectif de ce TP est de merger les bases welfare EHCVM du sénégal des années 2018 et 2021. Nous ferons les traitements necessaires afin d'avoir un bon résultat.

Section 1 : Installation des packages et Importation des bases

Installation des packages nécesssaires

```
packages <-  c("dplyr","haven", "gtsummary", "tidyverse", "labelled")

for (package in packages) {
  if(!requireNamespace(package,quietly = TRUE)){
    install.packages(package)
  }
  library(package,character.only = TRUE)
}
```

Importation des bases

```
welfare_2018 <- read_dta("../Donnees/ehcvm_welfare_sen2018.dta")
welfare_2021 <- read_dta("../Donnees/ehcvm_welfare_sen2021.dta")
```

Section 2: Traitement des bases

Description des bases

```
dim(welfare_2018)
```

```
## [1] 7156   35
```

```
colnames(welfare_2018)
```

```
##  [1] "country"   "year"      "hhid"      "grappe"    "menage"    "vague"
##  [7] "zae"       "region"    "milieu"    "hhweight"  "hhsize"    "eqadu1"
## [13] "eqadu2"    "hgender"   "hage"      "hmstat"    "hreligion" "hnation"
## [19] "halfab"    "heduc"     "hdiploma"  "hhandig"   "hactiv7j"  "hactiv12m"
## [25] "hbranch"   "hsectins"  "hcsp"      "dali"      "dnal"      "dtot"
## [31] "pcexp"     "zzae"      "zref"      "def_spa"   "def_temp"
```

L'analyse de ce résultat montre que la base welfare 2018 contient 7156 observations et 35 variables

```
dim(welfare_2021)
```

```
## [1] 7120   47
```

```
colnames(welfare_2021)
```

```
##  [1] "grappe"              "menage"              "country"
##  [4] "year"                "hhid"                "vague"
##  [7] "month"               "zae"                 "region"
## [10] "milieu"              "hhweight"            "hhsize"
## [13] "eqadu1"              "eqadu2"              "hgender"
## [16] "hage"                "hmstat"              "hreligion"
## [19] "hnation"             "hethnie"             "halfa"
## [22] "halfa2"              "heduc"               "hdiploma"
## [25] "hhandig"             "hactiv7j"            "hactiv12m"
## [28] "hbranch"             "hsectins"            "hcsp"
## [31] "dali"                "dnal"                "dtot"
## [34] "pcexp"               "zzae"                "zref"
## [37] "def_spa"             "def_temp"            "def_temp_prix2021m11"
## [40] "def_temp_cpi"        "def_temp_adj"        "zali0"
## [43] "dtet"                "monthly_cpi"         "cpi2017"
## [46] "icp2017"             "dollars"
```

La base welfare 2021 quant à elle contient 7120 observations et 47 variables.

Vérification des doublons

```
doublon_2018 <- welfare_2018[duplicated(welfare_2018[, c("grappe","menage")]), ] # sélectionne les do

doublon_2021 <- welfare_2021[duplicated(welfare_2021[, c("grappe","menage")]), ] # sélectionne les do

cat("Nombre de doublons présents dans la base welfare 2018  :", nrow(doublon_2018), "\n")
```

```
## Nombre de doublons présents dans la base welfare 2018  : 0
```

```
cat("Nombre de doublons présents dans la base welfare 2021  :", nrow(doublon_2021))
```

```
## Nombre de doublons présents dans la base welfare 2021  : 0
```

On constate qu'il n'y a pas de doublons dans les deux bases.

Vérification des valeurs manquantes

```
print("Valeurs manquantes base 2018")
```

```
## [1] "Valeurs manquantes base 2018"
```

```
NA_2018 <- colSums(is.na(welfare_2018))
NA_2018
```

```
##   country      year      hhid    grappe    menage     vague       zae    region
##         0         0         0         0         0         0         0         0
##    milieu  hhweight    hhsize    eqadu1    eqadu2   hgender      hage    hmstat
##         0         0         0         0         0         0         0         2
## hreligion   hnation    halfab     heduc  hdiploma   hhandig  hactiv7j hactiv12m
##         0         0         0         0         0         0         0         0
##   hbranch  hsectins      hcsp      dali      dnal      dtot     pcexp      zzae
##      1722      1722      1722         0         0         0         0         0
##      zref   def_spa  def_temp
##         0         0         0
```

```r
print("Valeurs manquante base 2021")
```

```
## [1] "Valeurs manquante base 2021"
```

```r
NA_2021 <- colSums(is.na(welfare_2021))
NA_2021
```

```
##               grappe               menage              country
##                    0                    0                    0
##                 year                 hhid                vague
##                    0                    0                    0
##                month                  zae               region
##                    0                    0                    0
##               milieu            hhweight               hhsize
##                    0                    0                    0
##               eqadu1              eqadu2              hgender
##                    0                    0                    0
##                 hage               hmstat            hreligion
##                    0                    0                    0
##              hnation              hethnie                halfa
##                    0                   82                    0
##               halfa2                heduc              hdiploma
##                    0                    0                    0
##              hhandig            hactiv7j            hactiv12m
##                    0                    0                    0
##              hbranch            hsectins                 hcsp
##                 1838                 1359                 1326
##                 dali                 dnal                 dtot
##                    0                    0                    0
##                pcexp                 zzae                 zref
##                    0                    0                    0
##              def_spa             def_temp def_temp_prix2021m11
##                    0                    0                    0
##         def_temp_cpi         def_temp_adj                 zali0
##                    0                    0                    0
##                 dtet          monthly_cpi              cpi2017
##                    0                    0                    0
##              icp2017              dollars
##                    0                    0
```

En analysant ce résultat, on constate que dans la base 2018, 4 variables présentent des valeurs manquantes dont 2 pour hmstat et 1722 pour les variables hbranch, hsectins et hcsp. Par contre sauf la variable hetnie possède 2 valeurs manquantes dans la base 2021.

Ces deux bases contiennent déjà la variable date pour distinguer les observation après fusion.

En outre, on constate une différence de variables entre les deux bases de données. Certaines variables sont présentes dans la base de 2021 mais absentes dans celle de 2018, et vice versa. De plus, certaines variables communes sont écrites différemment. Nous allons donc identifier ces variables et procéder aux corrections nécessaires.

Détection des variables identiques

Ici nous recherchons à avoir la liste des vairbales qui sont dans les deux bases

```r
# Trouver les noms de colonnes communes
variables_communes <- intersect(names(welfare_2021), names(welfare_2018))

# Affichage sous forme de chaîne de caractères
resultat <- paste0('"', variables_communes, '"', collapse = ", ")

cat(resultat, "\n")
```

```
## "grappe", "menage", "country", "year", "hhid", "vague", "zae", "region", "milieu", "hhweight", "hh
```

Les différences dans les noms des variables

```r
# Variables présentes uniquement en 2018
cat("Variables présentes uniquement en 2018 :\n")
```

```
## Variables présentes uniquement en 2018 :
```

```r
unique_2018 <- setdiff(names(welfare_2018), names(welfare_2021))
  print(unique_2018)
```

```
## [1] "halfab"
```

```r
# Variables présentes uniquement en 2021
cat("\nVariables présentes uniquement en 2021 :\n")
```

```
##
## Variables présentes uniquement en 2021 :
```

```r
unique_2021 <- setdiff(names(welfare_2021), names(welfare_2018))
  print(unique_2021)
```

```
##  [1] "month"                "hethnie"              "halfa"
##  [4] "halfa2"               "def_temp_prix2021m11" "def_temp_cpi"
##  [7] "def_temp_adj"         "zali0"                "dtet"
## [10] "monthly_cpi"          "cpi2017"              "icp2017"
## [13] "dollars"
```

On voit ainsi la différence entre les deux bases.

Dans la base 2018 la variable d'alphabetisation est nommé halfab alors que dans la base 2021 cette variable est décomposé en halfa Alphabétisation du CM (lire et écrire) et halfa2 Alphabétisation du CM (lire, écrire et comprend). En 2018, toute personne sachant lire et écrire dans une langue donnée est considérée comme alphabétisée dans cette langue. Donc nous allons renommer la variable halfab de l'EHCVM 2018 en halfa.

```r
welfare_2018 <- welfare_2018 %>%
  rename(halfa = halfab)
# Ajoutons cette variables à la liste des variables communes.

variables_communes <-  append(variables_communes,"halfab")
```

Dans la base 2021 nous avons la variable month qui n'est pas dans celle de 2018. Cette variable donne le jour, le mois et l'année de l'enquête vu que l'enquête s'est déroulée de 2021 à 2022. De même, les variables hethnie (éthnie), def_temp_prix2021m11 (déflateur temporel pour la pauvreté internationale), def_temp_cpi (déflateur temporel alternatif basé sur l'IPC officiel, style 2018/19), def_temp_adj (temporal deflator adjusted for difference between hh and market survey periods), zali0 (sum conso cp val up) variable créee, dtet (dépense totale annuelle par tête) variable créee, monthly_cpi (Valeur mensuelle de l'IPC),

cpi2017 (adjustment factor for inflation between 2017 ICP year and base period for survey), icp2017 (PPP exchange rate to USD based on 2017 ICP), dollars (welfare in 2017 PPP USD per capita per day (not spatially deflated)).

Les différences au niveau des modalités de chaque variable

Nous vérifions dans cetet parite si les modalités des variables sont les mêmes

```
variable_label_diff <- c() #créer une liste vide

for (variable in variables_communes) { #parcourir les variables en communs dans les deux bases

  if(labelled::is.labelled(welfare_2018[[variable]])){ #vérifier si la variable en 2018 est labéliser

      value_label_2018 <- labelled::val_labels(welfare_2018[[variable]]) #recupérer les labels de la

  }else{
      value_label_2018 <- NULL #Mettre vide dans le cas ou la variable en 2018 n'est pas labéliser
  }

  if(labelled::is.labelled(welfare_2021[[variable]])){ #vérifier si la variable en 2021 est labéliser

      value_label_2021 <- labelled::val_labels(welfare_2021[[variable]]) #recupérer les labels de la

  }else{

      value_label_2021 <- NULL #Mettre vide dans le cas ou la variable en 2021 n'est pas labéliser
  }

  if(!identical(value_label_2018, value_label_2021)){ #Vérifier si les labels de la variable  sont id

     variable_label_diff <- append(variable_label_diff,variable) #Si les labels diffèrent, alors ajou
     print(variable)

   }
}
```

```
## [1] "zae"
## [1] "hnation"
## [1] "hdiploma"
## [1] "hactiv7j"
## [1] "hbranch"
## [1] "hcsp"
```

En analysant ces sorties on constate que les variables zae, hnation, hdiploma, activité 7 jours, hbranch et hCSP sont labelisées différemment. Nous allons donc harmoniser cela.

Correction de la variable zae

```
labelled::val_labels(welfare_2018$zae)
```

```
## NULL
```

	Zone agroecologique						
	1	2	3	4	5	6	Total
Region residence							
dakar	1,020	0	0	0	0	0	1,020
ziguinchor	0	0	0	0	480	0	480
diourbel	0	0	552	0	0	0	552
SAINT-LOUIS	0	504	0	0	0	0	504
tambacounda	0	0	0	0	432	0	432
kaolack	0	0	0	528	0	0	528
thies	0	0	570	0	0	0	570
louga	0	0	480	0	0	0	480
fatick	0	0	0	455	0	0	455
kolda	0	0	0	0	432	0	432
matam	0	408	0	0	0	0	408
kaffrine	0	0	0	431	0	0	431
kedougou	0	0	0	0	0	456	456
sedhiou	0	0	0	0	408	0	408
Total	1,020	912	1,602	1,414	1,752	456	7,156

En 2018 la variable zae n'est pas labelisée

```
labelled::val_labels(welfare_2021$zae)
```

```
##                          Kédougou                Saint-Louis-Matam
##                                 1                                3
##            Thies-Diourbel-Louga         Kaolack-Fatick-Kaffrine
##                                 5                                7
## Ziguinchor-Tamba-Kolda-Sédhiou                            Dakar
##                                 9                               11
```

On constate qu'en 2021, cette variable est labelisée suivant un regroupement de région. Donc pour avoir les noms correspondants dans la base 2018, nous allons tabuler la variable zae et region de 2018.

```
welfare_2018 %>%labelled::to_factor()%>%
  tbl_cross(
    row = region,
    col = zae
  )
```

A partir de ce tableau on peut voir que le 1 correspond à Dakar, le 2 à Sant-louis et Matam, le 3 à Diourbel, Thies et Louga, le 4 à Kaloack, Fatick et Kaffrine, le 5 à Ziguinchor, Tambacounda et Louga et le 6 à kedougou. Pour une meilleure harmonisation, nous allons d'abord labeliser la variable zae dans la base 2018 et ensuite faire un recodange dans la base 2021.

```
welfare_2018 <- welfare_2018 %>%
  mutate(zae = labelled(zae,
    c("Dakar" = 1,
      "Saint-Louis-Matam" = 2,
      "Thies-Diourbel-Louga" = 3,
      "Kaolack-Fatick-Kaffrine" = 4,
      "Ziguinchor-Tamba-Kolda-Sédhiou" = 5,
```

```
      "Kédougou" = 6)
  ))
```

```
class(welfare_2021$zae)
```

```
## [1] "haven_labelled" "vctrs_vctr"     "double"
```

```
welfare_2021 <- welfare_2021 %>%
  mutate(zae = recode(as.factor(zae),  # Convertir en numérique pour éviter les erreurs
    `1` = 6,
    `3` = 2,
    `5` = 3,
    `7` = 4,
    `9` = 5,
    `11` = 1
  ))

#Affectons les labels de 2021 à ceux de 2018
labelled::val_labels(welfare_2021$zae) <-  labelled::val_labels(welfare_2018$zae)
# convertir en facteur
#welfare_2021 <- welfare_2021 %>%
#  mutate(zae = factor(zae))
```

Correction de la variable hnation

```
labelled::val_labels(welfare_2018$hnation)
```

```
##                   Benin            Burkina Faso           Côte d'Ivoire
##                       1                       2                       3
##           Guinée Bissau                    Mali                   Niger
##                       4                       5                       6
##                 Sénégal                    Togo                 Nigéria
##                       7                       8                       9
##            Autre CEDEAO           Autre Afrique Autre pays hors Afrique
##                      10                      11                      12
```

```
labelled::val_labels(welfare_2021$hnation)
```

```
##                   Bénin            Burkina Faso               Cape-vert
##                       1                       2                       3
##           Cote d'ivoire                  Gambie                   Ghana
##                       4                       5                       6
##                  Guinee           Guinée Bissau                 Liberia
##                       7                       8                       9
##                    Mali                   Niger                 Nigeria
##                      10                      11                      12
##                 Sénégal            Serra-Leonne                    Togo
##                      13                      14                      15
##           Autre Afrique Autre pays hors Afrique
##                      17                      18
```

On voit que dans les deux bases les pays associés à chaque numéro ne sont pas le même. Nous allons donc corriger cela

```
welfare_2021 %>%
  to_factor() %>% #labéliser
```

Répartition suivant les pays en 2021 avant correction	**N = 7,120**[1]
Nationalite du CM	
Bénin	0 (0%)
Burkina Faso	0 (0%)
Cape-vert	0 (0%)
Cote d'ivoire	1 (<0.1%)
Gambie	2 (<0.1%)
Ghana	1 (<0.1%)
Guinee	39 (0.5%)
Guinée Bissau	8 (0.1%)
Liberia	0 (0%)
Mali	18 (0.3%)
Niger	2 (<0.1%)
Nigeria	1 (<0.1%)
Sénégal	7,038 (99%)
Serra-Leonne	0 (0%)
Togo	1 (<0.1%)
Autre Afrique	7 (<0.1%)
Autre pays hors Afrique	2 (<0.1%)
Valeurs manquantes	0

[1] n (%)

```
  select(hnation) %>% #selection des variables qui vont s'afficher
  tbl_summary(digits = list(all_continuous() ~ 2), missing = "always", #afficher les valeurs manquant
              missing_text = "Valeurs manquantes")%>%
  modify_header(label = "Répartition suivant les pays en 2021 avant correction")
```

```
welfare_2021 <- welfare_2021 %>%
  mutate(hnation = recode(as.factor(hnation),  # Convertir en numérique si nécessaire
                          `4` = 3,
                          `8` = 4,
                          `10` = 5,
                          `11` = 6,
                          `13` = 7,
                          `15` = 8,
                          `12` = 9,
                          `17` = 11,
                          `18` = 12,
                          `3` = 10, `5` = 10, `6` = 10, `7` = 10, `9` = 10, `14` = 10
                          ))
```

Affectons les labels de 2018 à ceux de 2021

```
labelled::val_labels(welfare_2021$hnation) <-  labelled::val_labels(welfare_2018$hnation)
```

```
welfare_2021 %>%
  to_factor() %>% #labéliser
  select(hnation) %>% #selection des variables qui vont s'afficher
```

Répartition suivant les pays en 2021 après correction	**N = 7,120**[1]
hnation	
Benin	0 (0%)
Burkina Faso	0 (0%)
Côte d'Ivoire	1 (<0.1%)
Guinée Bissau	8 (0.1%)
Mali	18 (0.3%)
Niger	2 (<0.1%)
Sénégal	7,038 (99%)
Togo	1 (<0.1%)
Nigéria	1 (<0.1%)
Autre CEDEAO	42 (0.6%)
Autre Afrique	7 (<0.1%)
Autre pays hors Afrique	2 (<0.1%)
Valeurs manquantes	0

[1] n (%)

```
  tbl_summary(missing = "always", #afficher les valeurs manquantes
              missing_text = "Valeurs manquantes")%>%
  modify_header(label = "Répartition suivant les pays en 2021 après correction")
```

Correction de la variable hdiploma

Detection de l'incohérence

```
val_labels(welfare_2018$hdiploma)
```

```
##            Aucun         CEP/CFEE        BEPC/BFEM              cap               bt
##               0                1                2                3                4
##              bac  DEUG, DUT, BTS          Licence         Maitrise Master/DEA/DESS
##               5                6                7                8                9
##    Doctorat/Phd
##              10
```

```
val_labels(welfare_2021$hdiploma)
```

```
##            Aucun             cepe             bepc              cap               bt
##               0                1                2                3                4
##              bac  DEUG, DUT, BTS          Licence         Maitrise Master/DEA/DESS
##               5                6                7                8                9
##    Doctorat/Phd
##              10
```

On voit une différence dans l'écriture de bepc. Nous allons dans la suite corriger cela

```
welfare_2018 %>%
  to_factor() %>%
  select(hdiploma) %>%
  tbl_summary(missing = "always",
              missing_text = "NA")%>%
```

Répartition suivant les diplomes en 2018 avant correction	**N = 7,156**[1]
Diplome du CM	
Aucun	5,697 (80%)
CEP/CFEE	587 (8.2%)
BEPC/BFEM	359 (5.0%)
cap	52 (0.7%)
bt	16 (0.2%)
bac	154 (2.2%)
DEUG, DUT, BTS	49 (0.7%)
Licence	83 (1.2%)
Maitrise	63 (0.9%)
Master/DEA/DESS	65 (0.9%)
Doctorat/Phd	31 (0.4%)
NA	0

[1] n (%)

Répartition suivant les diplomes en 2018 après correction	**N = 7,156**[1]
Diplome du CM	
Aucun	5,697 (80%)
cepe	587 (8.2%)
bepc	359 (5.0%)
cap	52 (0.7%)
bt	16 (0.2%)
bac	154 (2.2%)
DEUG, DUT, BTS	49 (0.7%)
Licence	83 (1.2%)
Maitrise	63 (0.9%)
Master/DEA/DESS	65 (0.9%)
Doctorat/Phd	31 (0.4%)
NA	0

[1] n (%)

```
  modify_header(label = "Répartition suivant les diplomes en 2018 avant correction")
```

```
val_labels(welfare_2018$hdiploma) <-  val_labels(welfare_2021$hdiploma)
```

```
welfare_2018 %>%
  to_factor() %>%
  select(hdiploma) %>%
  tbl_summary(missing = "always",
              missing_text = "NA")%>%
  modify_header(label = "Répartition suivant les diplomes en 2018 après correction")
```

Répartition suivant l'activité du CM en 2018 avant correction	**N = 7,156**[1]
Activite 7 jours du CM	
Occupe	5,362 (75%)
Chomeur	44 (0.6%)
TF cherchant emploi	3 (<0.1%)
TF cherchant pas	60 (0.8%)
Inactif	1,687 (24%)
Moins de 5 ans	0 (0%)
NA	0

[1] n (%)

Correction de la variable hactivty7j

```
val_labels(welfare_2018$hactiv7j)
```

```
##              Occupe             Chomeur TF cherchant emploi    TF cherchant pas
##                   1                   2                   3                   4
##             Inactif      Moins de 5 ans
##                   5                   6
```

```
val_labels(welfare_2021$hactiv7j)
```

```
##              Occupe TF cherchant emploi    TF cherchant pas             Chomeur
##                   1                   2                   3                   4
##             Inactif      Moins de 5 ans
##                   5                   6
```

Ici on voit par exemple que chomeur est associe à 2 dans la base 2018 alors qu'il est associé à 4 dans la base 2021. Nous allons harmoniser cela

```
welfare_2018 %>%
  to_factor() %>%
  select(hactiv7j) %>%
  tbl_summary(missing = "always",
              missing_text = "NA")%>%
  modify_header(label = "Répartition suivant l'activité du CM en 2018 avant correction")
```

```
welfare_2018 <- welfare_2018 %>%
  mutate(hactiv7j = dplyr::recode(hactiv7j,
                           `2` = 4,
                           `3` = 2,
                           `4` = 3))
```

```
val_labels(welfare_2018$hactiv7j) <-  val_labels(welfare_2021$hactiv7j)
```

```
welfare_2018 %>%
  to_factor() %>%
  select(hactiv7j) %>%
  tbl_summary(missing = "always",
              missing_text = "NA")%>%
  modify_header(label = "Répartition suivant l'activité du CM en 2018 après correction")
```

Répartition suivant l'activité du CM en 2018 après correction	**N = 7,156**[1]
Activite 7 jours du CM	
Occupe	5,362 (75%)
TF cherchant emploi	3 (<0.1%)
TF cherchant pas	60 (0.8%)
Chomeur	44 (0.6%)
Inactif	1,687 (24%)
Moins de 5 ans	0 (0%)
NA	0

[1] n (%)

Correction de la variable branche d'activité

```
val_labels(welfare_2018$hbranch)
```

```
##      Agriculture    Elevage/peche    Indust. extr.    Autr. indust.
##                1                2                3                4
##              btp         Commerce Restaurant/Hotel     Trans./Comm.
##                5                6                7                8
##  Education/Sante  Services perso.    Aut. services
##                9               10               11
```

```
val_labels(welfare_2021$hbranch)
```

```
##        Agriculture Elevage/syl./peche      Indust. extr.      Autr. indust.
##                  1                  2                  3                  4
##                btp           Commerce   Restaurant/Hotel       Trans./Comm.
##                  5                  6                  7                  8
##    Education/Sante    Services perso.      Aut. services
##                  9                 10                 11
```

Ici on voit une différence dans les labels. En 2018, c'est Elevage/peche qui est associé à 2 alors qu'en 2021 c'est Elevage/syl./peche. Nous allons adopter la labelisation de 2021

```
val_labels(welfare_2018$hbranch) <-  val_labels(welfare_2021$hbranch)
```

Correction de la variable CSP

```
val_labels(welfare_2018$hcsp)
```

```
##                                        Cadre supérieur
##                                                      1
##                             Cadre moyen/agent de maîtrise
##                                                      2
##                               Ouvrier ou employé qualifié
##                                                      3
##                           Ouvrier ou employé non qualifié
##                                                      4
##                                  Manœuvre, aide ménagère
##                                                      5
##                            Stagiaire ou Apprenti rénuméré
```

```
##                                                          6
##                            Stagiaire ou Apprenti non rénuméré
##                                                          7
## Travailleur familial contribuant à une entreprise familiale
##                                                          8
##                                 Travailleur pour compte propre
##                                                          9
##                                                      Patron
##                                                         10
```

```
val_labels(welfare_2021$hcsp)
```

```
##                                               Cadre supérieur
##                                                             1
##                                   Cadre moyen/agent de maîtrise
##                                                             2
##                                     Ouvrier ou employé qualifié
##                                                             3
##                                   Ouvrier ou employé non qualifié
##                                                             4
##                                         Manœuvre, aide ménagère
##                                                             5
##                                    Stagiaire ou Apprenti rénuméré
##                                                             6
##                                Stagiaire ou Apprenti non rénuméré
##                                                             7
## Travailleur Familial contribuant pour une entreprise familial
##                                                             8
##                                    Travailleur pour compte propre
##                                                             9
##                                                         Patron
##                                                            10
```

Ici au niveau de la modalité 9, familial est écrit avec un f minuscule en 2018 et avec majuscule en 2021. Nous allons adopter la labelisation de 2018

```
val_labels(welfare_2021$hcsp) <-  val_labels(welfare_2018$hcsp)
```

Vérification après correction

```
variable_label_diff <- c() #créer une liste vide

for (variable in variables_communes) { #parcourir les variables en communs dans les deux bases

  if(labelled::is.labelled(welfare_2018[[variable]])){ #vérifier si la variable en 2018 est labéliser

      value_label_2018 <- labelled::val_labels(welfare_2018[[variable]]) #recupérer les labels de la

  }else{
      value_label_2018 <- NULL #Mettre vide dans le cas ou la variable en 2018 n'est pas labéliser
  }

  if(labelled::is.labelled(welfare_2021[[variable]])){ #vérifier si la variable en 2021 est labéliser
```

```
      value_label_2021 <- labelled::val_labels(welfare_2021[[variable]]) #recupérer les labels de la

  }else{

      value_label_2021 <- NULL #Mettre vide dans le cas ou la variable en 2021 n'est pas labéliser
  }

  if(!identical(value_label_2018, value_label_2021)){ #Vérifier si les labels de la variable  sont id

     variable_label_diff <- append(variable_label_diff,variable) #Si les labels diffèrent, alors ajou
     print(variable)

   }
}
```

Ainsi toutes les correction ont été bien faites.

Vérification des types de variables

Dans cette partie après correction des différences nous verrons nos variables communes ont le même type

```
# Initialiser un vecteur vide pour stocker les variables avec des types différents
variable_type_diff <- vector("character")

# Boucle sur les variables communes
for (variable in variables_communes) {

  # Récupérer les classes des variables en 2018 et 2021
  type_var_2018 <- class(welfare_2018[[variable]])
  type_var_2021 <- class(welfare_2021[[variable]])

  # Vérifier si les types sont différents
  if (!identical(type_var_2018, type_var_2021)) {
    variable_type_diff <- c(variable_type_diff, variable)  # Ajouter la variable au vecteur
  }
}

# Afficher les variables ayant des types différents entre 2018 et 2021
print(variable_type_diff)
```

```
## character(0)
```

Ainsi toutes les variables de même nom communes aux deux bases ont le même types.

Détection des ménages communes aux deux enquêtes

Un rapport publié par l'INS du Bénin mentionne que, dans la base **s00_me_ben2021**, la variable `PanelHH` indique si un ménage enquêté en 2021 avait déjà été interrogé en 2018. De même, cette variable est présente dans la base **s00_me_sen2021**, ce qui permet d'identifier les ménages communs aux deux enquêtes. Ainsi, nous allons fusionner ces deux bases, afin de repérer les ménages ayant participé aux enquêtes de 2018 et 2021.

	Alpha. lire/ecr./comp. CM		
	Non	Oui	Total
Alpha. lire/ecr. CM			
Non	3,477	0	3,477
Oui	121	3,522	3,643
Total	3,598	3,522	7,120

Importation de la base

```
Base_s00_me_sen2021 <-  read_dta("../Donnees/s00_me_sen2021.dta")
```

Fusion des deux bases

Nous allons effectuer la jointure en sélectionnant uniquement les variables `PanelHH`, `grappe` et `menage` dans la base **s00_me_ben2021**. Les variables `grappe` et `menage` seront utilisées comme clés de jointure afin d'identifier les ménages communs aux deux bases.

```
welfare_2021 <- dplyr::left_join(
  welfare_2021,
  Base_s00_me_sen2021 %>% select(grappe, menage, PanelHH),
  by = c("grappe", "menage")
)
```

Section 3: Fusion des bases welfare 2021 et 2018

Après avoir détecté et corrigé toutes les incohérences, nous allons, dans cette section, fusionner les deux bases en les **empilant**.

```
welfare_final <- bind_rows(welfare_2018,welfare_2021)
```

On note qu'après le merge il y a de nouvelles valeurs manquantes qui sont créées. Cela dû aux variables qui étaient présentes uniquement dans l'une des bases. Nous différencierons ces valeurs manquantes aux autres.

Par ailleurs, une tabulation entre les variables halfa et halfa2 de la base 2021 montre que tous les individus qui ont non pour halfa ont également non pour halfa2. Nous attribuerons ainsi la valeur non à ces individus pour la variable halfa2 dans la base finale.

Correction de la variable halfa

```
welfare_2021 %>%labelled::to_factor()%>%
  tbl_cross(
    row = halfa,
    col = halfa2
  )
```

Il nous faut avoir la description de la variable

```r
labelled::val_labels(welfare_2018$halfa)
```

```
## Non Oui
##   0   1
```

```r
welfare_final <- welfare_final %>%
  mutate(halfa2 = ifelse(halfa == 0 & year == 2018, 0, halfa2))
```

Enregistrement de la base

```r
write_dta(welfare_final, "../Sortie/welfare_final.dta")
```

Visualisation de la base finale

```r
# Sélectionner les 5 premières et 5 dernières variables
selected_vars <- c(names(welfare_final)[1:5], tail(names(welfare_final), 5))

# Afficher les 8 premières observations
head(welfare_final[selected_vars], 8)
```

```
## # A tibble: 8 x 10
##   country  year  hhid grappe menage monthly_cpi cpi2017 icp2017 dollars PanelHH
##   <chr>   <dbl> <dbl>  <dbl>  <dbl>       <dbl>   <dbl>   <dbl>   <dbl> <dbl+lb>
## 1 SEN      2018  1001      1      1          NA      NA      NA      NA NA
## 2 SEN      2018  1002      1      2          NA      NA      NA      NA NA
## 3 SEN      2018  1003      1      3          NA      NA      NA      NA NA
## 4 SEN      2018  2001      2      1          NA      NA      NA      NA NA
## 5 SEN      2018  2002      2      2          NA      NA      NA      NA NA
## 6 SEN      2018  2003      2      3          NA      NA      NA      NA NA
## 7 SEN      2018  2004      2      4          NA      NA      NA      NA NA
## 8 SEN      2018  2005      2      5          NA      NA      NA      NA NA
```

```r
# Afficher les 8 dernières observations
tail(welfare_final[selected_vars], 8)
```

```
## # A tibble: 8 x 10
##   country  year  hhid grappe menage monthly_cpi cpi2017 icp2017 dollars PanelHH
##   <chr>   <dbl> <dbl>  <dbl>  <dbl>       <dbl>   <dbl>   <dbl>   <dbl> <dbl+lb>
## 1 SEN      2021 59802    598      2        123.    1.10    239.    5.17 1 [Mena~
## 2 SEN      2021 59803    598      3        123.    1.10    239.    4.36 1 [Mena~
## 3 SEN      2021 59801    598      1        123.    1.10    239.    5.29 1 [Mena~
## 4 SEN      2021 59806    598      6        123.    1.10    239.    9.08 1 [Mena~
## 5 SEN      2021 59809    598      9        123.    1.10    239.    2.51 1 [Mena~
## 6 SEN      2021 59805    598      5        123.    1.10    239.    3.06 1 [Mena~
## 7 SEN      2021 59804    598      4        123.    1.10    239.    5.25 1 [Mena~
## 8 SEN      2021 59808    598      8        123.    1.10    239.    3.69 1 [Mena~
```