TP9 : Fusion des bases welfare de l’EHCVM 2018 et 2021

KAFANDO G. Judicaël Oscar

Table of Contents

[**Section 0 2**](#_Toc194253003)

[**Installation et importation des packages 2**](#_Toc194253004)

[**Chargement des bases 2**](#_Toc194253005)

[**Section 1 : Analyse et comparaison des bases** 2](#_Toc194253006)

[**Vérification des doublons** 2](#_Toc194253007)

[**Vérification des valeurs manquantes** 3](#_Toc194253008)

[**Informations collectées** 4](#_Toc194253009)

[**Vérification de la labélisation** 7](#_Toc194253010)

[**1. Variable “zae”** 8](#_Toc194253011)

[**2. Variable “hnation”** 10](#_Toc194253012)

[**3. Variable “hdiploma”** 13](#_Toc194253013)

[**4. Variable “hactiv7j”** 14](#_Toc194253014)

[**5. Variable “hbranch”** 16](#_Toc194253015)

[**6. Variable “hcsp”** 17](#_Toc194253016)

[**7. Vérifions finales des labels** 18](#_Toc194253017)

[**Vérification des types de variables** 19](#_Toc194253018)

[**Détection des ménages panels ou pas** 19](#_Toc194253019)

[**Section 2 : Fusion des bases** 20](#_Toc194253020)

[**Empilement** 20](#_Toc194253021)

[**Section 3 : Quelques statistiques** 21](#_Toc194253022)

# 

# Section 0

## Installation et importation des packages

Dans cette section, nous installons tous les packages qui servirons dans la suite.

packages <- c("readr","haven","utils","dplyr","labelled","questionr","gtsummary","ggplot2")  
  
  
for (package in packages) {  
 if (!requireNamespace(package, quietly = TRUE)) { # Vérifie si le package n'est pas encore installé  
 install.packages(package)  
 }  
 library(package, character.only = TRUE) # nom du package en nom ou chaine de caractère ()  
}

## Chargement des bases

## Base welfare  
  
welfare\_2018 <- haven::read\_dta("../Data/ehcvm\_welfare\_sen2018.dta") #importation de la base welfare 2018  
  
welfare\_2021 <- haven::read\_dta("../Data/ehcvm\_welfare\_sen2021.dta") #importation de la base welfare 2021

# **Section 1 : Analyse et comparaison des bases**

Pour un meilleur traitement, nous commençons par analyser et comprendre les deux bases, afin de voir les similitudes et les divergences.

## **Vérification des doublons**

doublon\_2018 <- welfare\_2018[duplicated(welfare\_2018[, c("grappe","menage")]), ] # sélectionne tous les ménages dupliqué qui sont dans la base 2018  
  
doublon\_2021 <- welfare\_2021[duplicated(welfare\_2021[, c("grappe","menage")]), ] # sélectionne tous les ménages dupliqué qui sont dans la base 2021  
  
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

Touts les deux bases n’ont pas de ménage dupliqué.

## **Vérification des valeurs manquantes**

# Calculer le nombre de valeurs manquantes par variable (colonne)  
print("Valeurs manquante base 2018")

## [1] "Valeurs manquante base 2018"

NA\_base\_2018 <- colSums(is.na(welfare\_2018))  
NA\_base\_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

print("Valeurs manquante base 2021")

## [1] "Valeurs manquante base 2021"

NA\_base\_2021 <- colSums(is.na(welfare\_2021))  
NA\_base\_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 considérant les deux bases, nous avons pas trop de valeurs manquantes. Seulement quatre variables en présentent en 2021 et en 2018.

## **Informations collectées**

Nous cherchons à avoir un aperçu de l’ensemble des variables issue des deux bases. Pour chacune, nous dirons si elle est dans une seule base ou dans les deux.

vars\_2018 = colnames(welfare\_2018) #liste des variables en 2018  
  
vars\_2021 = colnames(welfare\_2021) #liste des variables en 2021  
  
# Identifions les variables communes  
vars\_communes <- intersect(vars\_2018, vars\_2021) # prend les variables qui sont à la fois dans les deux bases  
  
# Variables présentes uniquement en 2018  
vars\_only\_2018 <- setdiff(vars\_2018, vars\_2021) #Garde les variables de vars\_2018 qui ne sont pas dans vars\_2021  
  
# Variables présentes uniquement en 2021  
vars\_only\_2021 <- setdiff(vars\_2021, vars\_2018) #Garde les variables de vars\_2021 qui ne sont pas dans vars\_2018

**Tableau récapitulatif**

# Créer un tableau récapitulatif   
recap <- tibble( #créons une mini base ayant la listes des variables  
 Variable = unique(c(vars\_2018, vars\_2021)), # garde les variables de façon unique, présente au moins dna sune des bases  
 `Présent en 2018` = ifelse(Variable %in% vars\_2018, "✅", ""), # Cocher vert si la variable est en 2018   
 `Présent en 2021` = ifelse(Variable %in% vars\_2021, "✅", "") #Cocher vert si la variable est en 2021  
)  
  
# Afficher le tableau avec gtsummary  
recap %>%  
 gt::gt() %>%  
 gt::tab\_header(  
 title = "Comparaison des variables entre Welfare 2018 et 2021",  
 subtitle = "Variables communes et spécifiques à chaque année"  
 )

Table 1: Comparaison des variables entre Welfare 2018 et 2021

Variables communes et spécifiques à chaque année

| Variable | Présent en 2018 | Présent en 2021 |
| --- | --- | --- |
| country | ✅ | ✅ |
| year | ✅ | ✅ |
| hhid | ✅ | ✅ |
| grappe | ✅ | ✅ |
| menage | ✅ | ✅ |
| vague | ✅ | ✅ |
| zae | ✅ | ✅ |
| region | ✅ | ✅ |
| milieu | ✅ | ✅ |
| hhweight | ✅ | ✅ |
| hhsize | ✅ | ✅ |
| eqadu1 | ✅ | ✅ |
| eqadu2 | ✅ | ✅ |
| hgender | ✅ | ✅ |
| hage | ✅ | ✅ |
| hmstat | ✅ | ✅ |
| hreligion | ✅ | ✅ |
| hnation | ✅ | ✅ |
| halfab | ✅ |  |
| heduc | ✅ | ✅ |
| hdiploma | ✅ | ✅ |
| hhandig | ✅ | ✅ |
| hactiv7j | ✅ | ✅ |
| hactiv12m | ✅ | ✅ |
| hbranch | ✅ | ✅ |
| hsectins | ✅ | ✅ |
| hcsp | ✅ | ✅ |
| dali | ✅ | ✅ |
| dnal | ✅ | ✅ |
| dtot | ✅ | ✅ |
| pcexp | ✅ | ✅ |
| zzae | ✅ | ✅ |
| zref | ✅ | ✅ |
| def\_spa | ✅ | ✅ |
| def\_temp | ✅ | ✅ |
| month |  | ✅ |
| hethnie |  | ✅ |
| halfa |  | ✅ |
| halfa2 |  | ✅ |
| def\_temp\_prix2021m11 |  | ✅ |
| def\_temp\_cpi |  | ✅ |
| def\_temp\_adj |  | ✅ |
| zali0 |  | ✅ |
| dtet |  | ✅ |
| monthly\_cpi |  | ✅ |
| cpi2017 |  | ✅ |
| icp2017 |  | ✅ |
| dollars |  | ✅ |

La base welfare de 2018 a **35 variables sur 7156 individus**, tandis que celle de 2021 a **47 sur 7120 individus**. L’on constate que de façon globale, les deux bases ont,**48 variables distinctes**. Parmi celles-ci, **34** sont présentes dans les deux bases, **13** seulement en 2021 et **1** seulement en 2018.

**NB : La variable halfab de 2018 et les variables halfa de 2021 donne la même information à savoir “Savoir lire et ecrire du CM”. Mais dans le tableau, du fait de l’orthographe, ils sont différents.**

Pour se faire nous renommons celle de 2021 en halfab.

# Renommons halfa en halfab  
  
colnames(welfare\_2021)[colnames(welfare\_2021)=="halfa"] <- "halfab"  
  
# Ajoutons cette variables à la liste des variables communes.  
  
vars\_communes <- append(vars\_communes,"halfab")

## **Vérification de la labélisation**

Dans cette section, nous verifions si pour les variables communes, nous avions les mêmes labels. Pour se faire, nous parcourons la liste des variables communes et nous récupérons les labels de chaque variable en 2018 et en 2021. Une fois cela fait, nous les comparons pour voir s’il y aincohérence ou pas.

variable\_label\_diff <- c() #créer une liste vide   
  
for (variable in vars\_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 variable en 2018  
  
 }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 variable en 2021  
   
 }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 identiques entre 2018 et 2021  
   
 variable\_label\_diff <- append(variable\_label\_diff,variable) #Si les labels diffèrent, alors ajouter le nom de la variable à la liste créer  
 print(variable)  
   
 }  
}

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

La taille de la liste contenant les variables dont les labels sont différents d’une année à l’autre est égale à 6. Donc nous avons **6 variable** qui ne sont pas labéliser de la meme façon. Traitons les cas par cas afin d’avoir leur labélisation identique.

### **1. Variable “zae”**

Elle représente la zone agré-écologique. Cette variable n’est pas labéliser en 2018 mais l’est en 2021. Donc pour uniformiser, nous labelisons en 2018 en utilisant les memes labels qu’en 2021.

**modalités de la variable en 2018**

labelled::val\_labels(welfare\_2018$zae)

## NULL

**modalités de la variable en 2021**

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

**Recodons la variable en 2021**

Par ce processus, nous aurons les memes codifications en 2021 et 2018 et après nous mettrons les labels sur la variable de 2018. Nous constatons que chaque *zone agré-écologique* est liée à la région de résidnce. En se basant sur cette variable, on arrive à recoder de façon conforme à celle de la *zone agré-écologique* de la base 2018.

**Distribution avant modification**

welfare\_2021 %>%  
 to\_factor() %>% #labéliser  
 select(zae) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Zone agroecologique |  |
| Kédougou | 452 (6.3%) |
| Saint-Louis-Matam | 911 (13%) |
| Thies-Diourbel-Louga | 1,599 (22%) |
| Kaolack-Fatick-Kaffrine | 1,413 (20%) |
| Ziguinchor-Tamba-Kolda-Sédhiou | 1,740 (24%) |
| Dakar | 1,005 (14%) |
| Valeurs manquantes | 0 |
| *1*n (%) | |

**Recodage**

welfare\_2021 <- welfare\_2021 %>%  
 dplyr::mutate(zae = dplyr::recode(zae,  
 `11` = 1, # Remplace la modalité 11 par 1  
 `9` = 5, # Remplace la modalité 9 par 5  
 `5` = 3, # Remplace la modalité 5 par 3  
 `3` = 2, # Remplace la modalité 3 par 2  
 `7` = 4, # Remplace la modalité 7 par 4  
 `1` = 6)) # Remplace la modalité 1 par 6

**Labélisation de 2021**

Le recodage n’a pas modifier la labélisation, donne nous l’adoptons à présent.

labelled::val\_labels(welfare\_2021$zae) <- c("Kédougou" = 6, "Saint-Louis-Matam"= 2, "Thies-Diourbel-Louga" = 3, "Kaolack-Fatick-Kaffrine" = 4, "Ziguinchor-Tamba-Kolda-Sédhiou" = 5,"Dakar" = 1)

**Distribution après modification**

welfare\_2021 %>%  
 to\_factor() %>% #labéliser  
 select(zae) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Zone agroecologique |  |
| Kédougou | 452 (6.3%) |
| Saint-Louis-Matam | 911 (13%) |
| Thies-Diourbel-Louga | 1,599 (22%) |
| Kaolack-Fatick-Kaffrine | 1,413 (20%) |
| Ziguinchor-Tamba-Kolda-Sédhiou | 1,740 (24%) |
| Dakar | 1,005 (14%) |
| Valeurs manquantes | 0 |
| *1*n (%) | |

Nous voyons que la structure est resté inchangé. Donc la labélisation a été bonne.

**Ajout des labels à 2018**

Maintenant que la codification est identique, nous Ajoutons à présent les labels qu’il faut à la base 2018

labelled::val\_labels(welfare\_2018$zae) <- labelled::val\_labels(welfare\_2021$zae) #Affectons juste les labels de 2021 à la base de 2018

### **2. Variable “hnation”**

**modalités de la variable en 2018**

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

**modalités de la variable en 2021**

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 note la présence de certains pays en 2021 qui n’y sont pas en 2018. il s’agit de *liberia, Serra-Leonne, Cap vert,Gambie, Gana et Guinée*, que nous mettrons dans la variable *Autre CEDEAO*. En plus de cela, nous réaménageons les modalités pour que ça concorde dans les deux bases.

**Distribution avant modification**

welfare\_2021 %>%  
 to\_factor() %>% #labéliser  
 select(hnation) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

| **Characteristic** | **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 (%) | |

**Modification (recodage)**

welfare\_2021 <- welfare\_2021 %>%  
 mutate(hnation = dplyr::recode(hnation,  
 `4` = 3, # Remplace la modalité 4 par 3  
 `8` = 4, # Remplace la modalité 8 par 4  
 `10` = 5, # Remplace la modalité 10 par 5  
 `11` = 6, # Remplace la modalité 11 par 6  
 `13` = 7, # Remplace la modalité 13 par 7  
 `15` = 8, # Remplace la modalité 15 par 8  
 `12` = 9, # Remplace la modalité 12 par 9  
 `17` = 11, # Remplace la modalité 17 par 11  
 `18` = 12, # Remplace la modalité 18 par 12  
 `3` = 10, `5` = 10, `6` = 10, `7` = 10, `9` = 10, `14` = 10 # Remplace les modalités (3,9,14,5,6,7) par 10  
 ))

**Ajout des labels**

Maintenant que les codes sont uniformiser, nous ajoutons les labels.

labelled::val\_labels(welfare\_2021$hnation) <- labelled::val\_labels(welfare\_2018$hnation) #Labélisons la variable hnation de 2018 en utilisant les labels de celle de 2018

**Distribution après modification**

welfare\_2021 %>%  
 to\_factor() %>% #labéliser  
 select(hnation) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Nationalite du CM |  |
| 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 (%) | |

### **3. Variable “hdiploma”**

**modalités de la variable en 2018**

labelled::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

**modalités de la variable en 2021**

labelled::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

Ici, l’incohérence a noté est uniquement au niveau de certains labels. Par exemple **cepe en 2021 et CEP/CFEE en 2018**. Sinon les modalités (codes) sont tous pareil. Pour se faire, nous conserverons les labélisations de 2018 et nous mettrons à jours celles de 2021 en fonction.

labelled::val\_labels(welfare\_2021$hdiploma) <- labelled::val\_labels(welfare\_2018$hdiploma)

### **4. Variable “hactiv7j”**

**modalités de la variable en 2018**

labelled::val\_labels(welfare\_2018$hactiv7j)

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

**modalités de la variable en 2021**

labelled::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, l’incohérence a noter est que les modalités et labels diffèrent peu. Par exmple en 2018 **2 representent chomeur** tandis qu’en 2021, **chomeur est representer par 4**.Pour se faire, nous conserverons les labélisations de 2018 et nous mettrons à jours celles de 2021 en fonction.

**Distribution avant modification**

welfare\_2021 %>%  
 to\_factor() %>% #labéliser  
 select(hactiv7j) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Activite 7 jours du CM |  |
| Occupe | 5,178 (73%) |
| TF cherchant emploi | 5 (<0.1%) |
| TF cherchant pas | 62 (0.9%) |
| Chomeur | 34 (0.5%) |
| Inactif | 1,841 (26%) |
| Moins de 5 ans | 0 (0%) |
| Valeurs manquantes | 0 |
| *1*n (%) | |

**Recodons celle de 2021**

Il s’agira juste de modifier les codes des modalités **chomeurs (4)**, **TF cherchant pas (3)** et **TF cherchant emploi (2)**

welfare\_2021 <- welfare\_2021 %>%  
 dplyr::mutate(hactiv7j = dplyr::recode(hactiv7j,  
 `4` = 2, # Remplace la modalité 4 par 2  
 `2` = 3, # Remplace la modalité 2 par 3  
 `3` = 4)) # Remplace la modalité 3 par 4

**Labélisation de la variable en 2021**

Maintenant que les codes sont bien agencés, nous affectons les labels de 2018 à 2021.

labelled::val\_labels(welfare\_2021$hactiv7j) <- labelled::val\_labels(welfare\_2018$hactiv7j)

**Distribution après modification**

welfare\_2021 %>%  
 to\_factor() %>% #labéliser  
 select(hactiv7j) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Activite 7 jours du CM |  |
| Occupe | 5,178 (73%) |
| Chomeur | 34 (0.5%) |
| TF cherchant emploi | 5 (<0.1%) |
| TF cherchant pas | 62 (0.9%) |
| Inactif | 1,841 (26%) |
| Moins de 5 ans | 0 (0%) |
| Valeurs manquantes | 0 |
| *1*n (%) | |

L’on constate que la structure reste inchangée

### **5. Variable “hbranch”**

**modalités de la variable en 2018**

labelled::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

**modalités de la variable en 2021**

labelled::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, l’incohérence a noter est uniquement au niveau de certains labels. Par exemple **Elevage/syl./peche en 2021 et Elevage/peche en 2018**. Sinon les modalités (codes) sont tous pareil. Pour se faire, nous conserverons les labélisations de 2021 et nous mettrons à jours celles de 2018 en fonction.

labelled::val\_labels(welfare\_2018$hbranch) <- labelled::val\_labels(welfare\_2021$hbranch)

### **6. Variable “hcsp”**

**modalités de la variable en 2018**

labelled::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

**modalités de la variable en 2021**

labelled::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, l’incohérence a noter est uniquement au niveau de certains labels. Par exemple **Travailleur familial contribuant pour une entreprise familiale en 2021 et Travailleur familial contribuant à une entreprise familiale en 2018**. Sinon les modalités (codes) sont tous pareil. Pour se faire, nous conserverons les labélisations de 2021 et nous mettrons à jours celles de 2018 en fonction.

labelled::val\_labels(welfare\_2018$hcsp) <- labelled::val\_labels(welfare\_2021$hcsp)

## **7. Vérifions finales des labels**

variable\_label\_diff <- c() #créer une liste vide   
  
for (variable in vars\_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 variable en 2018  
  
 }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 variable en 2021  
   
 }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 identiques entre 2018 et 2021  
   
 variable\_label\_diff <- append(variable\_label\_diff,variable) #Si les labels diffèrent, alors ajouter le nom de la variable à la liste créer  
 print(variable)  
   
 }  
}

Maintenant, La taille de la liste contenant les variables dont les labels sont différents d’une année à l’autre est égale à 0. Donc nous avons résolus ce problème maintenant.

## **Vérification des types de variables**

Toujours en considérant les variables communes; il faudrait s’assurer que d’une année à l’autre, la meme variables garde son type. Pour se faire nous utiliserons la fonction de base *class()*

variable\_type\_diff <- c() # Créer une liste vide pour stocker les variables avec des types différents  
  
for (variable in vars\_communes) { # Parcourir les variables communes dans les deux bases  
  
 type\_var\_2018 <- class(welfare\_2018[[variable]]) # Récupérer le type de la variable en 2018  
 type\_var\_2021 <- class(welfare\_2021[[variable]]) # Récupérer le type de la variable en 2021  
  
 if (!identical(type\_var\_2018, type\_var\_2021)) { # Vérifier si les types sont différents  
 variable\_type\_diff <- append(variable\_type\_diff, variable) # Ajouter la variable à la liste  
 }  
}  
  
# Afficher les variables ayant des types différents entre 2018 et 2021  
print(variable\_type\_diff)

## NULL

Nous n’avons aucune variable dont le type varie de 2018 à 2021. Pour chacune des variables, le type est le même.

## **Détection des ménages panels ou pas**

A présent, nous essaierons de voir parmi les ménages de 2021, lesquels étaient présent ou pas en 2018, et aussi ceux de 2018, lesquels n’y sont plus en 2021.

Pour se faire, un extrait du document explicatif de l’EHCVM 2021 nous guide : *<Dans chaque fichier, les variables « grappe » (unité primaire de sondage) et « menage » (numéro du ménage) servent à identifier le ménage et procéder à l’appariement. Pour les fichiers ayant un autre niveau (individu,champ, etc.), il y a une ou des variables supplémentaires facilement identifiables dans les bases de données. De manière similaire, les variables « grappe » et « ménage » permettent à l’utilisateur de fusionner les données de cette vague avec la vague précédente* ***(EHCVM1)****. Cependant, les utilisateurs doivent garder à l’esprit que seuls les ménages qu’ont été marqués comme panel dans la variable « PanelHH » auront une valeur valide) pour être fusionnés avec la vague précédente.>*

Ce qu’il faut retenir, c’est que pour chaque ménage de 2021, une variable montre si elle y était en 2018 ou pas. Cette variable est **PanelHH** est est présente dans la base identification du ménage (s00\_me\_sen\_2021 ). Allons donc chercher cette variable et ajoutons la à notre base welfare

**1. Importation et selection de la variable**

var\_panel <- haven::read\_dta("../Data/s00\_me\_sen2021.dta") %>% # Importation de la base  
 select(grappe,menage,PanelHH) # Sélection de la variable et des identifiants du ménage (grappe, id)

**2. Ajout de la variable à la base welfare**

Pour se faire, nous ferons une jointure en prenant la grappe et le numéro de chaque ménage comme identifiant

welfare\_2021\_ajusté <- dplyr::left\_join(welfare\_2021,var\_panel,by=c("grappe","menage"))

# **Section 2 : Fusion des bases**

Dans cette partie, nous proposerons une fusion, la plus adéquate possible à nos deux bases. La fusion consistera en un **empilement**; donc nous empilerons les données de 2018 et 2021.

### **Empilement**

Apres avoir vérifier les différentes incohérences possibles, nous empilons les données en utilisant la fonction *bind\_rows*.

welfare\_total <- bind\_rows(welfare\_2018,welfare\_2021\_ajusté) %>% # empiler les deux bases   
 arrange(grappe,menage) # trié suivant la grappe et le numéro du ménage

Maintenant que l’empilement est fait, de nouveaux manquants se sont crées, cela due aux variables non communes.

Dans la base \*welfare total (fusionnée)**, toutes les valeurs manquants** qui s’y trouve dans des variables propres à l’année 2021 et qui ont pour année (**year**) 2018 sont généré après la fusion. C’est d’eux qu’on qualifie de nouveaux manquants.

**Labélisons**

welfare\_total <- welfare\_total %>%   
 labelled::to\_factor()

**Extrait de la base**

welfare\_total[1:10,1:10] #affichons les 10 premieres lignes et les 10 premières variables

## # A tibble: 10 × 10  
## country year hhid grappe menage vague zae region milieu hhweight  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <dbl>  
## 1 SEN 2018 1001 1 1 1 Dakar dakar Urbain 1750.  
## 2 SEN 2018 1002 1 2 1 Dakar dakar Urbain 1750.  
## 3 SEN 2018 1003 1 3 1 Dakar dakar Urbain 1750.  
## 4 SEN 2018 2001 2 1 2 Dakar dakar Urbain 266.  
## 5 SEN 2021 201 2 1 2 Dakar dakar Urbain 290.  
## 6 SEN 2018 2002 2 2 2 Dakar dakar Urbain 266.  
## 7 SEN 2018 2003 2 3 2 Dakar dakar Urbain 266.  
## 8 SEN 2021 203 2 3 2 Dakar dakar Urbain 386.  
## 9 SEN 2018 2004 2 4 2 Dakar dakar Urbain 266.  
## 10 SEN 2021 204 2 4 2 Dakar dakar Urbain 290.

# **Section 3 : Quelques statistiques**

Table1 <- welfare\_total %>%  
 select(year, hgender, hage, hhsize, heduc, milieu, pcexp) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(by = year, #grouper par années  
   
 label = list( #labélisons les différentes variables  
 hgender ~ "Sexe du chef de ménage",  
 hage ~ "Âge (moyenne)",  
 hhsize ~ "Taille du ménage",  
 heduc ~ "Niveau d'éducation",  
 milieu ~ "Milieu de résidence",  
 pcexp ~ "Dépense per capital (FCFA)"  
 ),  
   
 missing = "no", # Ne pas inclure les NA  
   
 statistic = list(  
 all\_continuous() ~ "{mean} ± {sd}", #afficher la moyenne et l'écart-type pour les variables continues  
 all\_categorical() ~ "{n} ({p}%)" #afficher l'effectif et le pourcentage pour les variables catégorielles  
 ),  
   
 digits = all\_continuous() ~ 1, #mettre tous les nombres des variable numériques continue avec un chiffre apres la virgule  
 ) %>%  
   
 bold\_labels()  
  
Table1 # Afficher la table

| **Characteristic** | **2018** N = 7,156*1* | **2021** N = 7,120*1* |
| --- | --- | --- |
| **Sexe du chef de ménage** |  |  |
| Masculin | 5,280 (74%) | 5,095 (72%) |
| Féminin | 1,876 (26%) | 2,025 (28%) |
| **Âge (moyenne)** | 51.5 ± 14.1 | 54.1 ± 13.7 |
| **Taille du ménage** | 9.2 ± 6.0 | 8.7 ± 5.3 |
| **Niveau d'éducation** |  |  |
| Aucun | 4,833 (68%) | 4,894 (69%) |
| Maternelle | 2 (<0.1%) | 1 (<0.1%) |
| Primaire | 1,057 (15%) | 1,053 (15%) |
| Second. gl 1 | 524 (7.3%) | 522 (7.3%) |
| Second. tech. 1 | 47 (0.7%) | 10 (0.1%) |
| Second. gl 2 | 277 (3.9%) | 268 (3.8%) |
| Second. tech. 2 | 18 (0.3%) | 27 (0.4%) |
| Postsecondaire | 53 (0.7%) | 61 (0.9%) |
| Superieur | 345 (4.8%) | 284 (4.0%) |
| **Milieu de résidence** |  |  |
| Urbain | 3,941 (55%) | 3,922 (55%) |
| Rural | 3,215 (45%) | 3,198 (45%) |
| **Dépense per capital (FCFA)** | 615,630.2 ± 624,819.1 | 621,198.4 ± 536,958.0 |
| *1*n (%); Mean ± SD | | |

Le tableau ci-dessous montre qu’entre 2018 et 2021, parmi les ménages enquetés, la proportion de **femmes chefs de ménage** a légèrement augmenté **(26% à 28%)**, tandis que celle des hommes a diminué **(74% à 72%)**. **L’âge moyen** des chefs de ménage est passé de **51,5 à 54,1 ans**, traduisant un vieillissement de cette population. La **taille moyenne des ménages** a diminué de **9,2 à 8,7 personnes**, indiquant une tendance vers des ménages plus réduits.

Concernant **l’éducation**, la proportion des chefs de ménage n’ayant **jamais été scolarisés a légèrement augmentée (68% à 69%)** . Le pourcentage ayant atteint **l’enseignement primaire ou secondaire** est resté stable, tandis que ceux ayant une **formation technique ou supérieure** ont légèrement diminué.

La répartition entre **milieu urbain (55%) et rural (45%)** est restée inchangée. Enfin, la **dépense annuelle moyenne par habitant a légèrement augmenté (615 630 FCFA en 2018 à 621 198 FCFA en 2021)**, avec toutefois une grande variabilité entre ménages.