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 (ENSAE | Pierre Ndiaye)

TP9: Merge des bases welfare 2018 et 2021

Rédigé par :

FOGWOUNG DJOUFACK Sarah-Laure

NIASS Ahmadou

NGUEMFOUO NGOUMTSA Célina

SENE Malick

Élèves Ingénieurs Statisticiens Économistes

Sous la supervision de :

M. Aboubacar HEMA

Research Analyst

Année scolaire : 2024/2025

Contents

CONSIGNE DU TP4

L'objectif de ce TP est de fusionner les bases de données welfare 2018 et welfare 2021 en assurant une harmonisation et vérification rigoureuse des correspondances des labels entre les deux bases.

Execution du travail à faire

1- Importation et chargement des packages

Dans cette étape, nous allons vérifier si les packages nécessaires sont installés, les installer si besoin, puis les charger. Pour charger un package, nous utilisons la fonction library() du package utils.

```
# Liste des packages nécessaires
packages <- c("readr", "haven",
              "utils", "dplyr", "gtsummary", "labelled")

# Fonction pour vérifier et installer un package manquant
installer_si_manquant <- function(package) {
  if (!requireNamespace(package, quietly = TRUE)) {
    install.packages(package, dependencies = TRUE)
  }
  # Chargement du package avec library() du package utils
  library(package, character.only = TRUE)
}

# Appliquer la fonction à tous les packages listés
lapply(packages, installer_si_manquant)
```

2- Importation des bases de données

Dans cette étape, nous allons importer les bases de données welfare de 2018 et 2021.Pour cela, nous utilisons la fonction read_dta du package haven.

```
# Chargement de la base de données 2018
welfare_2018 <- haven::read_dta(
  "../Donnees/ehcvm_welfare_sen2018.dta")

# Chargement de la base de données 2021
welfare_2021 <- haven::read_dta(
  "../Donnees/ehcvm_welfare_sen2021.dta")
```

3- Exploration des données

Dans cette étape, nous allons explorer les données en affichant un aperçu des premières lignes et les dimensions des bases.

Aperçu des premieres lignes des bases de données

```
# Affichage des premières lignes de chaque base
head(welfare_2018)
head(welfare_2021)
```

Dimension des bases

```
# Vérifier le nombre de lignes et de colonnes pour chaque base

dim_2018 <- dim(welfare_2018)
dim_2021 <- dim(welfare_2021)
dim_2018
```

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

```
dim_2021
```

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

On remarque ici que la base welfare de 2018 a porté sur 7156 individus et il y avait 35 variables tandis que pour celle de 2021 il y avait 7120 individus et elle portait sur 47 variables.

Structure des variables

```
# Vérifier la structure et le type des variables

glimpse(welfare_2018)
glimpse(welfare_2021)
```

4- Comparaison des variables entre les deux bases

Dans cette étape, nous allons identifier les variables communes aux deux bases et vérifier que leurs types et labels sont identiques ou pas.

Identification des variables communes aux deux bases.

```
## Lister toutes les variables des deux bases
vars_2018 <- colnames(welfare_2018)
vars_2021 <- colnames(welfare_2021)
# Identification des variables communes
vars_communes <- intersect(vars_2018, vars_2021)
print(vars_communes)
```

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

Il y a 34 variables communes aux deux bases.

Regardons si ces variables sont de meme type pour les deux bases.

```
# Comparaison des types des variables communes dans les deux bases
types_communes_2018 <- vapply(welfare_2018[vars_communes],
                              function(x) paste(class(x),
                              collapse = " "), character(1))
types_communes_2021 <- vapply(welfare_2021[vars_communes],
                              function(x) paste(class(x),
                              collapse = " "), character(1))

# Création du tableau de comparaison
```

```r
comparaison <- data.frame(
  Variable = vars_communes,
  Type_2018 = types_communes_2018,
  Type_2021 = types_communes_2021,
  Same_Type = types_communes_2018 == types_communes_2021
)

# Affichage du résultat
print(comparaison)
```

```
##            Variable                          Type_2018
## country     country                          character
## year           year                            numeric
## hhid           hhid                            numeric
## grappe       grappe                            numeric
## menage       menage                            numeric
## vague         vague                            numeric
## zae             zae                            numeric
## region       region haven_labelled vctrs_vctr double
## milieu       milieu haven_labelled vctrs_vctr double
## hhweight   hhweight                            numeric
## hhsize       hhsize                            numeric
## eqadu1       eqadu1                            numeric
## eqadu2       eqadu2                            numeric
## hgender     hgender haven_labelled vctrs_vctr double
## hage           hage                            numeric
## hmstat       hmstat haven_labelled vctrs_vctr double
## hreligion hreligion haven_labelled vctrs_vctr double
## hnation     hnation haven_labelled vctrs_vctr double
## heduc         heduc haven_labelled vctrs_vctr double
## hdiploma   hdiploma haven_labelled vctrs_vctr double
## hhandig     hhandig haven_labelled vctrs_vctr double
## hactiv7j   hactiv7j haven_labelled vctrs_vctr double
## hactiv12m hactiv12m haven_labelled vctrs_vctr double
## hbranch     hbranch haven_labelled vctrs_vctr double
## hsectins   hsectins haven_labelled vctrs_vctr double
## hcsp           hcsp haven_labelled vctrs_vctr double
## dali           dali                            numeric
## dnal           dnal                            numeric
```

```
## dtot           dtot                                numeric
## pcexp         pcexp                                numeric
## zzae           zzae                                numeric
## zref           zref                                numeric
## def_spa     def_spa                                numeric
## def_temp   def_temp                                numeric
##                                          Type_2021 Same_Type
## country                                  character      TRUE
## year                                       numeric      TRUE
## hhid                                       numeric      TRUE
## grappe                                     numeric      TRUE
## menage                                     numeric      TRUE
## vague                                      numeric      TRUE
## zae       haven_labelled vctrs_vctr double     FALSE
## region    haven_labelled vctrs_vctr double      TRUE
## milieu    haven_labelled vctrs_vctr double      TRUE
## hhweight                                   numeric      TRUE
## hhsize                                     numeric      TRUE
## eqadu1                                     numeric      TRUE
## eqadu2                                     numeric      TRUE
## hgender   haven_labelled vctrs_vctr double      TRUE
## hage                                       numeric      TRUE
## hmstat    haven_labelled vctrs_vctr double      TRUE
## hreligion haven_labelled vctrs_vctr double      TRUE
## hnation   haven_labelled vctrs_vctr double      TRUE
## heduc     haven_labelled vctrs_vctr double      TRUE
## hdiploma  haven_labelled vctrs_vctr double      TRUE
## hhandig   haven_labelled vctrs_vctr double      TRUE
## hactiv7j  haven_labelled vctrs_vctr double      TRUE
## hactiv12m haven_labelled vctrs_vctr double      TRUE
## hbranch   haven_labelled vctrs_vctr double      TRUE
## hsectins  haven_labelled vctrs_vctr double      TRUE
## hcsp      haven_labelled vctrs_vctr double      TRUE
## dali                                       numeric      TRUE
## dnal                                       numeric      TRUE
## dtot                                       numeric      TRUE
## pcexp                                      numeric      TRUE
## zzae                                       numeric      TRUE
## zref                                       numeric      TRUE
```

```
## def_spa                                  numeric       TRUE
## def_temp                                 numeric       TRUE
```

On remarque que ces variables communes sont bien de meme type pour les deux bases et maintenant nous allons voir si les label pour les variables catégorielles communes sont identiques dans les deux bases sauf pour la variables zae zone agroécologique) qui dans la base 2018 est numeric et sans correspondances de labels pour chaque levels, tandis que dans la base 2021 elle est bien labellisé, de ce fait regardons d'abord si pour cette variable les valeurs prises dans les deux bases sont les memes.

```
## Regardonc les modalités de la variable zae dans les deux bases
unique(welfare_2018$zae)
```

```
## [1] 1 5 3 2 4 6
```

```
unique(welfare_2021$zae)
```

```
## <labelled<double>[6]>: Zone agroecologique
## [1] 11  9  5  3  7  1
##
## Labels:
##  value                          label
##      1                       Kédougou
##      3              Saint-Louis-Matam
##      5           Thies-Diourbel-Louga
##      7        Kaolack-Fatick-Kaffrine
##      9 Ziguinchor-Tamba-Kolda-Sédhiou
##     11                          Dakar
```

On remarque meme que tandis qu'en 2018, cette variable prend les valeurs 1,5,3,2,4 et 6; en 2021, les valeurs prises sont plutot 1,3, 5,7,9 et 11. Et donc on ne saurait faire de correspondances. Pour le moment, nous allons continuer le travail avec les autres variables communes.

Verification de la correspondance de labels

On va afficher les level+labels de toutes les variables communes catégorielles des deux bases. Puis on va identifier celles qui ont des level+labels differentes pour les variables communes.

```r
# Initialisation d'une liste pour stocker les différences de labels
differences_labels <- list()

# Parcours de chaque variable commune
for (var in vars_communes) {

  # Vérifions l'existence de la variable dans les deux bases
  if (var %in% colnames(welfare_2018) & var %in% colnames(welfare_2021)) {

    # Récupérons les labels avec la fonction val_labels()
    labels_2018 <- labelled::val_labels(welfare_2018[[var]])
    labels_2021 <- labelled::val_labels(welfare_2021[[var]])

    # Vérifions s'il y a des différences entre
    # les labels des deux années
    if (!identical(labels_2018, labels_2021)) {
      differences_labels[[var]] <- list("2018" = labels_2018,
                                        "2021" = labels_2021)
    }
  }
}

# Affichons les variables ayant des labels différents
cat("Variables ayant des labels différents entre 2018 et 2021 : ")
```

```
## Variables ayant des labels différents entre 2018 et 2021 :
```

```r
print(differences_labels)
```

```
## $zae
## $zae$`2018`
## NULL
##
## $zae$`2021`
##                       Kédougou             Saint-Louis-Matam
##                              1                             3
##         Thies-Diourbel-Louga      Kaolack-Fatick-Kaffrine
##                              5                             7
## Ziguinchor-Tamba-Kolda-Sédhiou                        Dakar
```

```
##                          9                         11
##
##
## $hnation
## $hnation$`2018`
##             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
##
## $hnation$`2021`
##             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
##
##
## $hdiploma
## $hdiploma$`2018`
##       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
##
## $hdiploma$`2021`
```

```
##       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
##
##
## $hactiv7j
## $hactiv7j$`2018`
##         Occupe         Chomeur TF cherchant emploi   TF cherchant pas
##             1              2              3              4
##             Inactif      Moins de 5 ans
##                  5                  6
##
## $hactiv7j$`2021`
##         Occupe TF cherchant emploi   TF cherchant pas         Chomeur
##             1              2              3              4
##             Inactif      Moins de 5 ans
##                  5                  6
##
##
## $hbranch
## $hbranch$`2018`
##    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
##
## $hbranch$`2021`
##     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
##
```

```
##
## $hcsp
## $hcsp$`2018`
##                                                 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
##
## $hcsp$`2021`
##                                                 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
```

On remarque une divergence de labels pour les variables hnation, hdiploma, hactiv7j, hbranch et hcsp (sachant que la variable zae est d'abord mise de coté car on ne saurait faire la correspondance).

5- Correction pour l'harmonisation des labels des variables communes

Dans cette étape, nous corrigeons les labels des variables divergentes pour les harmoniser entre 2018 et 2021.Nous allons procéder variable par variable.

Variable hnation:

Nous allons ajuster les codes de 2021 pour qu'ils correspondent à ceux de 2018. En effet, dans la base 2021, les autres pays de la CEDEAO ont été détaillés individuellement, tandis que dans la base 2018, ces pays étaient regroupés, donc c'est ce processus qui semble plus simple et judicieux. Par la suite, on verifie que les correspondances avec les effectifs sont bien maintenus.

En effet, Dans la base 2021, la variable hnation est définie avec 18 codes, où 1 = Bénin, 2 = Burkina Faso, 3 = Cape-vert, 4 = Cote d'ivoire, 5 = Gambie, 6 = Ghana, 7 = Guinee, 8 = Guinée Bissau, 9 = Liberia, 10 = Mali, 11 = Niger, 12 = Nigeria, 13 = Sénégal, 14 = Serra-Leonne, 15 = Togo, 17 = Autre Afrique et 18 = Autre pays hors Afrique.

En revanche, dans la base 2018, la variable hnation utilise 12 codes dont 10 apparaissent dans les données, avec 1 = Benin, 2 = Burkina Faso, 3 = Côte d'Ivoire, 4 = Guinée Bissau, 5 = Mali, 6 = Niger, 7 = Sénégal, 8 = Togo, 9 = Nigéria, 10 = Autre CEDEAO, 11 = Autre Afrique et 12 = Autre pays hors Afrique.

```
# Création de copies des bases pour modification
base2021_modifiee <- welfare_2021
base2018_modifiee <- welfare_2018
```

```r
# 1. Visualisation des fréquences initiales
#de la variable hnation dans la base 2021

# Ici, nous utilisons la fonction to_factor() du package
# haven pour convertir les valeurs en facteurs
# et observer les libellés.

freq_avant <- table(to_factor(welfare_2021$hnation))
print(freq_avant)
```

```
##
##              Bénin       Burkina Faso          Cape-vert
##                  0                  0                  0
##       Cote d'ivoire             Gambie              Ghana
##                  1                  2                  1
##             Guinee      Guinée Bissau            Liberia
##                 39                  8                  0
##              Mali              Niger            Nigeria
##                 18                  2                  1
##           Sénégal       Serra-Leonne               Togo
##              7038                  0                  1
##      Autre Afrique Autre pays hors Afrique
##                  7                  2
```

```r
#Correction des codes dans la base 2021 pour harmoniser avec 2018
base2021_modifiee <- base2021_modifiee %>%
  mutate(
    hnation = as.numeric(as.character(hnation))
  ) %>%
dplyr::mutate(
    hnation = dplyr::case_when(
      hnation == 1 ~ 1,    # Le level de Bénin est le meme
      hnation == 2 ~ 2,    # Le level de Burkina Faso est le meme
      # Cape-vert entre dans "Autre CEDEAO"
      hnation == 3 ~ 10,
      # Le 4 de la Cote d'Ivoire est plutot 3 dans la base 2018
      hnation == 4 ~ 3,
      # Gambie est mis dans "Autre CEDEAO"
      hnation == 5 ~ 10,
```

```
      # Ghana est intégré dans "Autre CEDEAO"
      hnation == 6 ~ 10,
      # Guinee est intégré dans "Autre CEDEAO"
      hnation == 7 ~ 10,
      # Le 8 de la Guinee Bissau est plutot 4 dans la base 2018
      hnation == 8 ~ 4,
      # Liberia est intégré dans "Autre CEDEAO"
      hnation == 9 ~ 10,
      # Le 10 de la Mali est plutot 5 dans la base 2018
      hnation == 10 ~ 5,
      # Le 11 du Niger est plutot 6 dans la base 2018
      hnation == 11 ~ 6,
      # Le 12 du Nigeria est plutot 9 dans la base 2018
      hnation == 12 ~ 9,
      # Le 13 du Senegal est plutot 7 dans la base 2018
      hnation == 13 ~ 7,
      # Serra-Leonne est intégré dans "Autre CEDEAO"
      hnation == 14 ~ 10,
      # Le 15 du Togo est plutot 8 dans la base 2018
      hnation == 15 ~ 8,
      # La correspondance 17 du Autre Afrique
      # est plutot 11 dans la base 2018
      hnation == 17 ~ 11,
      # Le 18 du Autre pays hors Afrique
      # est plutot 12 dans la base 2018
      hnation == 18 ~ 12,
      TRUE ~ NA_real_
    )
  )

# Attribution des labels pour la variable hnation dans la base 2021
#Ici, nous utilisons la fonction labelled() du package labelled
# pour redéfinir les libellés en fonction des codes harmonisés.
base2021_modifiee$hnation <- labelled(
  base2021_modifiee$hnation,
  c(
    "Benin" = 1,
    "Burkina Faso" = 2,
    "Côte d'Ivoire" = 3,
```

```
    "Guinée Bissau" = 4,
    "Mali" = 5,
    "Niger" = 6,
    "Sénégal" = 7,
    "Togo" = 8,
    "Nigéria" = 9,
    "Autre CEDEAO" = 10,
    "Autre Afrique" = 11,
    "Autre pays hors Afrique" = 12
  )
)

# 4. Vérification finale : affichage des fréquences après recodage
#pour confirmer la correspondance avec la base 2018

freq_apres <- table(to_factor(base2021_modifiee$hnation))
print(freq_apres)
```

```
##
##             Benin      Burkina Faso     Côte d'Ivoire
##                 0                 0                 1
##      Guinée Bissau              Mali             Niger
##                 8                18                 2
##           Sénégal              Togo           Nigéria
##              7038                 1                 1
##       Autre CEDEAO    Autre Afrique Autre pays hors Afrique
##                42                 7                 2
```

Après vérification, on constate que les effectifs de chaque modalité ont été conservés après harmonisation de labellisation.

Variable hactiv7j

Dans la base 2021, la variable hactiv7j présente une inversion dans la codification par rapport à la base 2018. Concrètement, dans la base 2018, le code 2 correspond à TF cherchant emploi et le code 4 à Chômeur, alors que dans la base 2021, ces codes sont inversés. Pour harmoniser les deux bases, nous allons recoder la variable hactiv7j de 2021 afin d'adapter sa structure à celle de 2018. Nous vérifierons ensuite que les effectifs de chaque modalité sont correctement conservés après recodage.

```r
# Visualisation initiale des fréquences de hactiv7j de 2021

freq_avant <- table(to_factor(welfare_2021$hactiv7j))
print(freq_avant)
```

```
##
##          Occupe TF cherchant emploi   TF cherchant pas          Chomeur
##            5178                   5                 62               34
##             Inactif      Moins de 5 ans
##                1841                   0
```

```r
# Correction: Recodage de hactiv7j dans la base 2021
# pour harmoniser la codification avec la base 2018.

# Recodage numérique selon correspondance 2021 → 2018
#   Code 1 reste 1 (Occupe)
#   Code 2 devient 3 (TF cherchant emploi)
#   Code 3 devient 4 (TF cherchant pas)
#   Code 4 devient 2 (Chômeur)
#   Codes 5 et 6 restent inchangés (Inactif et Moins de 5 ans)

base2021_modifiee <- base2021_modifiee %>%
  dplyr::mutate(
    hactiv7j = case_when(
      hactiv7j == 1 ~ 1,
      hactiv7j == 2 ~ 3,
      hactiv7j == 3 ~ 4,
      hactiv7j == 4 ~ 2,
      hactiv7j == 5 ~ 5,
      hactiv7j == 6 ~ 6,
      TRUE ~ NA_real_
    )
  )

# Attribution des labels pour hactiv7j dans la base 2021 modifiée,
# en s'assurant que la correspondance avec la base 2018 est respectée.

base2021_modifiee$hactiv7j <- labelled(
  base2021_modifiee$hactiv7j,
```

```
  c(
    "Occupe" = 1,
    "Chomeur" = 2,
    "TF cherchant emploi" = 3,
    "TF cherchant pas" = 4,
    "Inactif" = 5,
    "Moins de 5 ans" = 6
  )
)

# Vérification finale : affichage des fréquences après recodage
# pour confirmer la conservation des effectifs

freq_apres <- table(to_factor(base2021_modifiee$hactiv7j))
print(freq_apres)
```

```
##
##          Occupe         Chomeur TF cherchant emploi   TF cherchant pas
##            5178              34                   5                 62
##             Inactif      Moins de 5 ans
##                1841                   0
```

Variable hdiploma

Dans la base de 2021, les diplômes sont nommés différemment de ceux de 2018. Nous allons donc uniformiser les noms en harmonisant CEP/CFEE et BEPC/BFEM qui étaient écrits cepe et bepc respectivement dans la base 2021.

```
# Visualisation des fréquences avant modification
freq_avant <- table(to_factor(welfare_2021$hdiploma))
print(freq_avant)
```

```
##
##        Aucun         cepe         bepc          cap           bt
##         5772          583          317           39            7
##          bac  DEUG, DUT, BTS      Licence     Maitrise Master/DEA/DESS
##          150           46          101           55           32
##    Doctorat/Phd
##              18
```

```r
# CORRECTION DES LABELS

# 1. Extraction des labels actuels de la variable hdiploma
labels_avant <- attr(base2021_modifiee$hdiploma, "labels")

# 2. Identification des codes correspondant aux diplômes à renommer
code_cepe <- which(labels_avant == 1)
code_bepc <- which(labels_avant == 2)

# 3. Modification des labels sans changer leur level car correct
names(labels_avant)[code_cepe] <- "CEP/CFEE"
names(labels_avant)[code_bepc] <- "BEPC/BFEM"

# 4. Réassignation des labels mis à jour
attr(base2021_modifiee$hdiploma, "labels") <- labels_avant

# Vérification après modification
freq_apres <- table(to_factor(base2021_modifiee$hdiploma))
print(freq_apres)
```

```
##
##        Aucun      CEP/CFEE     BEPC/BFEM          cap           bt
##         5772          583          317           39            7
##          bac  DEUG, DUT, BTS      Licence     Maitrise Master/DEA/DESS
##          150           46          101           55           32
##    Doctorat/Phd
##              18
```

Après harmonisation, nous avons bien aligné les labels CEP/CFEE et BEPC/BFEM avec ceux de 2018, tout en conservant les effectifs de chaque modalité.

hbranch

Dans la base de 2021, le niveau 2 de hbranch inclut la sylviculture en plus de l'élevage et de la pêche (c'est Elevage/syl./peche). Pour assurer l'uniformité avec 2018, nous allons mettre à jour le libellé correspondant.

```r
# Visualisation des fréquences avant modification
freq_avant <- table(to_factor(welfare_2018$hbranch))
print(freq_avant)
```

```
##
##    Agriculture   Elevage/peche   Indust. extr.   Autr. indust.
##           1366             374              58             497
##            btp      Commerce Restaurant/Hotel    Trans./Comm.
##            313            1094              63             251
##  Education/Sante  Services perso.    Aut. services
##              379              761               278
```

```r
# CORRECTION DES LABELS (sans changer le type )

# Extraction des labels actuels de la variable hbranch
labels_avant <- attr(base2018_modifiee$hbranch, "labels")

# Identification du code correspondant au niveau 2
code_level2 <- which(labels_avant == 2)

# Modification du libellé en ajoutant la sylviculture
names(labels_avant)[code_level2] <- "Elevage/syl./peche"

# Réassignation des labels mis à jour
attr(base2018_modifiee$hbranch, "labels") <- labels_avant

# Vérification après modification
freq_apres <- table(to_factor(base2018_modifiee$hbranch))
print(freq_apres)
```

```
##
##     Agriculture Elevage/syl./peche    Indust. extr.    Autr. indust.
##            1366              374              58             497
##             btp         Commerce  Restaurant/Hotel     Trans./Comm.
##             313             1094              63             251
##    Education/Sante    Services perso.      Aut. services
##                379                761               278
```

Après cette mise à jour, nous avons bien intégré la sylviculture dans la modalité correspondante de 2018. Les effectifs des différentes catégories restent bien inchangés.

hcsp

Dans la base de 2021, l'intitulé du label correspondant à la valeur 8 est "Travailleur contribuant pour une entreprise familiale". Cependant, dans la base de 2018, la version correcte est "Travailleur contribuant à une entreprise familiale". Et comme en francais, la forme la plus correcte est "Travailleur contribuant à une entreprise familiale", nous modifions le libellé de 2021 pour qu'il corresponde à la version de 2018.

```
# Visualisation des fréquences avant correction
freq_avant <- table(to_factor(welfare_2021$hcsp))
print(freq_avant)
```

```
##
##                                                Cadre supérieur
##                                                             57
##                                  Cadre moyen/agent de maîtrise
##                                                            280
##                                    Ouvrier ou employé qualifié
##                                                            450
##                                Ouvrier ou employé non qualifié
##                                                            332
##                                       Manœuvre, aide ménagère
##                                                            151
##                                 Stagiaire ou Apprenti rénuméré
##                                                             34
##                             Stagiaire ou Apprenti non rénuméré
##                                                              3
## Travailleur Familial contribuant pour une entreprise familial
##                                                             66
##                                  Travailleur pour compte propre
##                                                           4302
##                                                         Patron
##                                                            119
```

```
# CORRECTION DES LABELS

# Extraction des labels actuels
labels_avant <- attr(base2021_modifiee$hcsp, "labels")
```

```r
# Identification du code à modifier
code_level8 <- which(labels_avant == 8)

# Modification du libellé pour la valeur 8
names(labels_avant)[code_level8] <-
  "Travailleur familial contribuant à une entreprise familiale"

# Réassignation des labels corrigés
attr(base2021_modifiee$hcsp, "labels") <- labels_avant

# Vérification après modification
freq_apres <- table(to_factor(base2021_modifiee$hcsp))
print(freq_apres)
```

```
##
##                                          Cadre supérieur
##                                                       57
##                                Cadre moyen/agent de maîtrise
##                                                      280
##                                  Ouvrier ou employé qualifié
##                                                      450
##                               Ouvrier ou employé non qualifié
##                                                      332
##                                       Manœuvre, aide ménagère
##                                                      151
##                                Stagiaire ou Apprenti rénuméré
##                                                       34
##                              Stagiaire ou Apprenti non rénuméré
##                                                        3
## Travailleur familial contribuant à une entreprise familiale
##                                                       66
##                                 Travailleur pour compte propre
##                                                     4302
##                                                   Patron
##                                                      119
```

L'harmonisation a bien été faite tout en conservant les effectifs de chaque modalité.

6- Vérification de la correspondance des labels entre les variables

Nous allons maintenant comparer les labels des variables harmonisées entre les bases de 2018 et 2021 afin de nous assurer qu'ils sont bien identiques après les corrections effectuées.

```
# Charger le package nécessaire
library(labelled)

# Fonction pour comparer les labels
compare_labels <- function(var1, var2, name) {
  # Labels de la variable dans la base 2018
  labels1 <- val_labels(var1)
  # Labels de la variable dans la base 2021
  labels2 <- val_labels(var2)
  # Vérifie si les labels sont identiques
  identical_labels <- identical(labels1, labels2)

  if (identical_labels) {
    cat(paste0("Les labels de ", name,
               " sont identiques dans les deux bases.\n"))
  } else {
    cat(paste0("Les labels de ", name, " sont différents.\n"))
    # Labels présents dans 2018 mais pas en 2021
    print(setdiff(names(labels1), names(labels2)))
    # Labels présents en 2021 mais pas en 2018
    print(setdiff(names(labels2), names(labels1)))
  }
}

# Comparer les labels des variables communes
compare_labels(base2018_modifiee$hcsp,
               base2021_modifiee$hcsp, "hcsp")
```

```
## Les labels de hcsp sont identiques dans les deux bases.
```

```
compare_labels(base2018_modifiee$hbranch,
               base2021_modifiee$hbranch, "hbranch")
```

```
## Les labels de hbranch sont identiques dans les deux bases.
```

```
compare_labels(base2018_modifiee$hdiploma,
               base2021_modifiee$hdiploma, "hdiploma")
```

```
## Les labels de hdiploma sont identiques dans les deux bases.
```

```
compare_labels(base2018_modifiee$hactiv7j,
               base2021_modifiee$hactiv7j, "hactiv7j")
```

```
## Les labels de hactiv7j sont identiques dans les deux bases.
```

```
compare_labels(base2018_modifiee$hnation,
               base2021_modifiee$hnation, "hnation")
```

```
## Les labels de hnation sont identiques dans les deux bases.
```

Après exécution de cette vérification, nous constatons que les labels sont bien harmonisés entre les deux bases. Nous pouvons donc passer à l'étape suivante.

7- Renommage des variables non communes aux deux bases

Afin de conserver une traçabilité des données spécifiques à chaque année, nous allons ajouter le suffixe correspondant à l'année (_2018 ou _2021) aux variables qui ne sont présentes que dans une seule des deux bases.

```
# 1. Identifions les variables spécifiques à chaque base

common_vars <- intersect(names(base2018_modifiee),
                         names(base2021_modifiee))
unique_2018  <- setdiff(names(base2018_modifiee), common_vars)
unique_2021  <- setdiff(names(base2021_modifiee), common_vars)

# Affichage des variables spécifiques à chaque base
print("\n les variables uniquement présentes en 2018")
```

```
## [1] "\n les variables uniquement présentes en 2018"
```

```r
print(unique_2018)
```

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

```r
print("\n les variables uniquement présentes en 2021")
```

```
## [1] "\n les variables uniquement présentes en 2021"
```

```r
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"
```

```r
# 2. Renommage des variables spécifiques à chaque année
base2018_modifiee <- base2018_modifiee %>%
  rename_with(~ paste0(.,"_2018"), all_of(unique_2018))

base2021_modifiee <- base2021_modifiee %>%
  rename_with(~ paste0(.,"_2021"), all_of(unique_2021))

# 3. Vérification du renommage
unique_2018_corr  <- setdiff(names(base2018_modifiee), common_vars)
unique_2021_corr  <- setdiff(names(base2021_modifiee), common_vars)
print("\n Vérication de la correction en 2018")
```

```
## [1] "\n Vérication de la correction en 2018"
```

```r
print(unique_2018_corr)
```

```
## [1] "halfab_2018"
```

```r
print("\n Vérication de la correction en 2021")
```

```
## [1] "\n Vérication de la correction en 2021"
```

```r
print(unique_2021_corr)
```

```
##  [1] "month_2021"                  "hethnie_2021"
##  [3] "halfa_2021"                  "halfa2_2021"
##  [5] "def_temp_prix2021m11_2021" "def_temp_cpi_2021"
##  [7] "def_temp_adj_2021"           "zali0_2021"
##  [9] "dtet_2021"                   "monthly_cpi_2021"
## [11] "cpi2017_2021"                "icp2017_2021"
## [13] "dollars_2021"
```

```r
# 4. Renommer aussi zae car la correspondance n'a pas été faite
# Et donc pour cette variables,
# nous distinguons zae pour l'année 2018 et pour 2021
base2018_modifiee <- base2018_modifiee %>%
  rename(zae_2018 = zae)

base2021_modifiee <- base2021_modifiee %>%
  rename(zae_2021 = zae)
```

8- Fusion des deux bases

Nous allons maintenant fusionner les bases de 2018 et 2021 en faisant un append. Il y avait 34 variables communes (mais on n'a pas pu faire la correspondance donc on la distinguera our l'année 2018 et pour 2021 et donc ici ca fait deja 35variables) et 1 variable uniquement présentes dans la base 2018 et 13 uniquement présentes en 2021, et donc dans la base finale, il doit avoir 35+1+13= 49 variables. Et par ailleurs 7120 (en 2018) + 7156 (en 2021) observations= 14276 observations

```r
# Append les deux bases
base_finale <- bind_rows(base2018_modifiee, base2021_modifiee)

# Vérifier la structure finale
print(dim(base_finale))
```

```
## [1] 14276    49
```

```
# Enregistrement
write_dta(base_finale,"../Sortie/base_finale.dta")
```

Les statistiques prevues pour la base finale sont verifiées donc c'est ok.