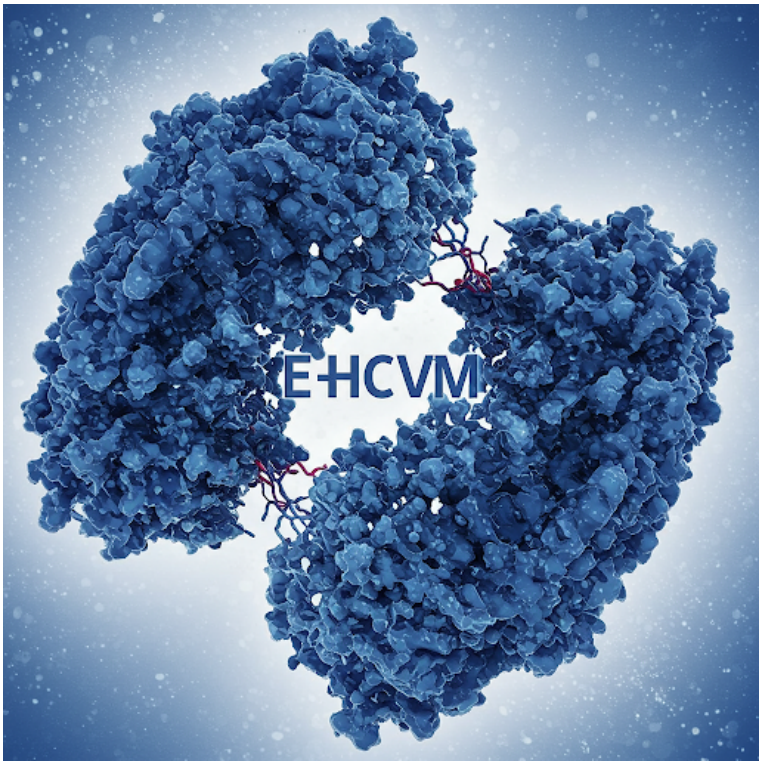
EHCVM 2018 et 2021

Aissatou Sega DIALLO

2025-03-20

Fusion des bases Welfare EHCVM



Sous la supervision de :
M. Hema

Introduction

Ce travail pratique (TP) a pour objectif de fusionner les bases de données welfare des Enquêtes sur les Conditions de Vie des Ménages (EHCVM) de 2018 et 2021 du Sénégal, afin de créer une base de données consolidée qui permettra de réaliser des analyses comparatives. Nous procéderons à l'exploration et à la préparation des données pour des analyses statistiques futures, en traitant des problématiques telles que les valeurs manquantes, les doublons et les valeurs aberrantes. À la fin de ce TP, la base de données fusionnée sera nettoyée et prête à être exploitée pour des études ultérieures sur les conditions de vie des ménages sénégalais.

I.Installation et chargement des packages

Cette section permet d'assurer que les bibliothèques requises sont correctement installées et chargées pour l'exécution des analyses.

```
libraries <- c("readr", "dplyr","labelled","tidyr", "gtsummary","haven", "utils", "ggplot2", "plotly")

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

II.Importation des bases de données welfare

Cette section consiste à importer les bases de données welfare des Enquêtes sur les Conditions de Vie des Ménages (EHCVM) afin de les préparer pour l'analyse.

```
getwd()
```

```
## [1] "C:/Formation ISE/ISE 1/Logiciel R/Projet-statistique-sous-R/TP 9/TP9_DIALLO_AissatouSega/Script
```

```
welf2018 <-  haven::read_dta("../Données/ehcvm_welfare_sen2018.dta") #importation base welfare 2018
welf2021 <-  haven::read_dta("../Données/ehcvm_welfare_sen2021.dta") #importation base welfare 2021
```

III.Exploration des données :

Cette section permettra d'examiner la structure des données et d'obtenir un aperçu des variables. Nous analyserons les premières lignes de chaque fichier afin de confirmer leur bonne importation et leur préparation pour l'analyse.

```
# Cette étape permet d'examiner rapidement la structure des données
head(welf2018)
```

```
## # A tibble: 6 x 35
##   country  year  hhid grappe menage vague   zae region   milieu  hhweight hhsize
```

```
##   <chr>   <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl+lb> <dbl+l>    <dbl>  <dbl>
## 1 SEN      2018  1001      1      1     1     1 1 [daka~ 1 [Urb~    1750.      2
## 2 SEN      2018  1002      1      2     1     1 1 [daka~ 1 [Urb~    1750.      2
## 3 SEN      2018  1003      1      3     1     1 1 [daka~ 1 [Urb~    1750.      1
## 4 SEN      2018  2001      2      1     2     1 1 [daka~ 1 [Urb~     266.     10
## 5 SEN      2018  2002      2      2     2     1 1 [daka~ 1 [Urb~     266.      6
## 6 SEN      2018  2003      2      3     2     1 1 [daka~ 1 [Urb~     266.      4
## # i 24 more variables: eqadu1 <dbl>, eqadu2 <dbl>, hgender <dbl+lbl>,
## #   hage <dbl>, hmstat <dbl+lbl>, hreligion <dbl+lbl>, hnation <dbl+lbl>,
## #   halfab <dbl+lbl>, heduc <dbl+lbl>, hdiploma <dbl+lbl>, hhandig <dbl+lbl>,
## #   hactiv7j <dbl+lbl>, hactiv12m <dbl+lbl>, hbranch <dbl+lbl>,
## #   hsectins <dbl+lbl>, hcsp <dbl+lbl>, dali <dbl>, dnal <dbl>, dtot <dbl>,
## #   pcexp <dbl>, zzae <dbl>, zref <dbl>, def_spa <dbl>, def_temp <dbl>
```

```
head(welf2021)
```

```
## # A tibble: 6 x 47
##   grappe menage country  year  hhid vague month      zae         region   milieu
##    <dbl>  <dbl> <chr>   <dbl> <dbl> <dbl> <date>     <dbl+lbl>  <dbl+lb> <dbl+l>
## 1      2      5 SEN      2021   205     2 2022-05-01 11 [Dakar] 1 [daka~ 1 [Urb~
## 2      2     15 SEN      2021   215     2 2022-05-01 11 [Dakar] 1 [daka~ 1 [Urb~
## 3      2      3 SEN      2021   203     2 2022-05-01 11 [Dakar] 1 [daka~ 1 [Urb~
## 4      2     13 SEN      2021   213     2 2022-05-01 11 [Dakar] 1 [daka~ 1 [Urb~
## 5      2      8 SEN      2021   208     2 2022-06-01 11 [Dakar] 1 [daka~ 1 [Urb~
## 6      2     16 SEN      2021   216     2 2022-06-01 11 [Dakar] 1 [daka~ 1 [Urb~
## # i 37 more variables: hhweight <dbl>, hhsize <dbl>, eqadu1 <dbl>,
## #   eqadu2 <dbl>, hgender <dbl+lbl>, hage <dbl>, hmstat <dbl+lbl>,
## #   hreligion <dbl+lbl>, hnation <dbl+lbl>, hethnie <dbl+lbl>, halfa <dbl+lbl>,
## #   halfa2 <dbl+lbl>, heduc <dbl+lbl>, hdiploma <dbl+lbl>, hhandig <dbl+lbl>,
## #   hactiv7j <dbl+lbl>, hactiv12m <dbl+lbl>, hbranch <dbl+lbl>,
## #   hsectins <dbl+lbl>, hcsp <dbl+lbl>, dali <dbl>, dnal <dbl>, dtot <dbl>,
## #   pcexp <dbl>, zzae <dbl>, zref <dbl>, def_spa <dbl>, def_temp <dbl>, ...
```

```
# Vérification des dimensions des deux bases de données
dim(welf2018)
```

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

```
dim(welf2021)
```

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

```
# Informations générales sur les variables/colonnes
str(welf2018)
str(welf2021)
```

IV. Évaluation de la qualité des données :

Dans cette section, l'analyse de la qualité des données sera faite en vérifiant la présence de valeurs manquantes, de doublons, et en s'assurant que les types de données correspondent bien aux variables attendues. Cette étape nous aidera à identifier d'éventuelles anomalies.

1. Doublons

```
# Vérification des doublons
sum(duplicated(welf2018))
```

```
## [1] 0
```

```
sum(duplicated(welf2021))
```

```
## [1] 0
```

Aucun doublon n'a été trouvé dans les deux bases

2. Valeurs manquantes

```
## Détection des valeurs manquantes
# Résumé des données welf2018
summary(welf2018)
# Résumé des données welf2021
summary(welf2021)
```

```
## Calcul du pourcentage de valeurs manquantes pour chaque colonne
colSums(is.na(welf2018)) / nrow(welf2018) * 100
colSums(is.na(welf2021)) / nrow(welf2021) * 100

# Un calcul du pourcentage de valeurs manquantes sera fait pour chaque colonne des deux bases. Le mode
```

Pour la base de données welfare 2018, les variables hactiv7j, hsectins, dnal, pcexp, hcsp, hbranch, hactiv12m, dtot, zzae et dali présentent des pourcentages relativement faibles de valeurs manquantes, variant entre 0 % et 24 %. Il en va de même pour la base welfare 2021, avec les variables hcsp, hbranch, dali, hethnie et hsectins.

V. Comparaison des variables et modalités entre les bases welfare 2018 et 2021

Après la visualisation les données et vérification de la présence d'éventuelles anomalies, un procédé à une comparaison des variables et des modalités entre les bases welfare 2018 et 2021 sera engagé. Cette étape nous permettra d'identifier les divergences dans l'annotation des variables et des modalités entre les deux bases.

```
#Liste des variables des deux bases
var2018=colnames(welf2018)
var2021=colnames(welf2021)

# Identification des variables communes aux deux bases
var_communes <- intersect(var2018, var2021)
```

```r
# Variables spécifiques à la base 2018
var_sp2018 <- setdiff(var2018, var2021)

# Variables spécifiques à la base 2021
var_sp2021 <- setdiff(var2021, var2018)

var_sp2018
```

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

```r
var_sp2021
```

```
##  [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"
```

Cette catégorisation montre que les deux bases partagent 34 variables communes. La base welfare 2018 comporte une seule variable unique, halfab, tandis que 13 variables sont spécifiques à la base welfare 2021 : “month”, “hethnie”, “halfa”, “halfa2”, “def_temp_prix2021m11”, “def_temp_cpi”, “def_temp_adj”, “zali0”, “dtet”, “monthly_cpi”, “cpi2017”, “icp2017” et “dollars”.

Concernant la variable d’alphabétisation, il s’agit simplement d’une différence d’orthographe. Nous procéderons donc à renommer la variable de 2021 en halfab et l’inclure dans les variables communes.

```r
## Renommage de la variable 'halfa' en 'halfab'
colnames(welf2021)[colnames(welf2021)=="halfa"] <- "halfab"

# Ajout de 'halfab' à la liste des variables communes
var_communes <- append(var_communes, "halfab")
```

Ainsi , a la suite , une vérification de la cohérence des labellisations.

1. Identification des incohérences de labellisation

```r
discord_lab_var <- c()

for (var in var_communes) {

  if(labelled::is.labelled(welf2018[[var]])){ # Vérification si la variable en 2018 est labellisée

      lab_val2018 <- labelled::val_labels(welf2018[[var]]) # Récupération des labels de la variable en 

  }else{
      lab_val2018 <- NULL # Si la variable en 2018 n'est pas labellisée, on l'indique par NULL
  }

  if(labelled::is.labelled(welf2021[[var]])){ # Vérification si la variable en 2021 est labellisée
```

```
      lab_val2021 <- labelled::val_labels(welf2021[[var]]) # Récupération des labels de la variable en

  }else{

      lab_val2021 <- NULL # Si la variable en 2021 n'est pas labellisée, on l'indique par NULL
  }

  if(!identical(lab_val2018, lab_val2021)){ # Vérification de la concordance des labels entre 2018 et 2

     discord_lab_var <- append(discord_lab_var,var) # Si les labels diffèrent, on ajoute le nom de la v
     print(var)

   }
}
```

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

Il a été constaté que six variables présentent des incohérences dans leur labellisation entre les bases welfare 2018 et 2021 : “zae”, “hnation”, “hdiploma”, “hactiv7j”, “hbranch” et “hcsp”. Afin de corriger ces divergences, chaque variable sera traitée individuellement, recodée et ses modalités ajustées en fonction des incohérences détectées.

2. Gestion des incohérences

a) Traitement de la variable hnation

```
# Visualisation de la distribution de la variable hnation en 2018
labelled::val_labels(welf2018$hnation)
```

Visualisation de la distribution

```
##                   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
```

```
# Visualisation de la distribution de la variable hnation en 2021
labelled::val_labels(welf2021$hnation)
```

Nationalité du Chef de Ménage	**N = 7,120**[1]
Nationalité 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 (%)

```
##                   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
```

```
# Transformation de la variable hnation en facteur et affichage de la distribution en 2021
welf2021 %>%
  to_factor() %>%
  select(hnation) %>%
  tbl_summary(
    missing = "always",  # Affichage des valeurs manquantes
    missing_text = "Valeurs manquantes",
    label = list(hnation ~ "Nationalité du CM"))%>%
    modify_header(label = "**Nationalité du Chef de Ménage**"
    )
```

Recherche de l'incoherence Certaines modalités, telles que "Autres CEDEAO" en 2018, ont été subdivisées en catégories distinctes, comme c'est le cas pour la Guinée, qui figure désormais comme une modalité spécifique.

Correction de l'incohérence Afin de corriger cette incohérence, la variable sera recodée dans la base welfare 2021, en combinant les modalités qui étaient regroupées sous "Autres CEDEAO" en 2018, y compris la Guinée et les autres pays concernés.

```
#Recodage dans la base de 2021
welf2021 <- welf2021 %>%
  mutate(hnation = dplyr::recode(hnation,
                           `4` = 3,  # Remplace la modalité 4 par 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
                           ))
```

```
#Uniformisation des labels
labelled::val_labels(welf2021$hnation) <-  labelled::val_labels(welf2018$hnation)
```

```
welf2021 %>%
  to_factor() %>%
  select(hnation) %>%
  tbl_summary(missing = "always",
              missing_text = "Valeurs manquantes",
              label = list(hnation ~ "Nationalité du CM"))%>%
    modify_header(label = "**Nationalité du Chef de Ménage**"
              )
```

Contrôle des modifications

b) Traitement de la variable hdiploma

```
# Affichage des labels de la variable hdiploma en 2018
labelled::val_labels(welf2018$hdiploma)
```

Visualisation de la distribution

```
##           Aucun        CEP/CFEE       BEPC/BFEM             cap              bt
##               0               1               2               3               4
```

Nationalité du Chef de Ménage	**N = 7,120**[1]
Nationalité 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 (%)

```
##             bac  DEUG, DUT, BTS         Licence        Maitrise Master/DEA/DESS
##               5               6               7               8               9
##    Doctorat/Phd
##              10
```

```
# Affichage des labels de la variable hdiploma en 2021
labelled::val_labels(welf2021$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
```

```
# Résumé statistique de la variable hdiploma en 2018 avec affichage des valeurs manquantes
welf2018 %>%
  to_factor() %>%
  select(hdiploma) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(hdiploma ~ "Diplôme du CM"))%>%
    modify_header(label = "**Diplôme le plus élevé du Chef de Ménage**"
              )
```

Recherche de l'incoherence Les différences observées entre les deux bases concernent uniquement des modifications d'étiquettes (labels) sans altération des valeurs sous-jacentes.

Correction de l'incohérence Il suffira donc d'affecter les labels de hdiploma dans welfare 2021 à ceux de welfare 2018 afin d'assurer une harmonisation des libellés entre les deux bases.

Diplôme le plus élevé du Chef de Ménage	**N = 7,156**[1]
Diplôme 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 (%)

```
#Harmonisation des labels
val_labels(welf2018$hdiploma) <-  val_labels(welf2021$hdiploma)
```

```
welf2018 %>%
  to_factor() %>%
  select(hdiploma) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(hdiploma ~ "Diplôme du CM"))%>%
    modify_header(label = "**Diplôme le plus élevé du Chef de Ménage**"
              )
```

Controle des modifications

c) Traitement de la variable hactiv7j

```
# Affichage des labels de la variable hactiv7j en 2018
labelled::val_labels(welf2018$hactiv7j)
```

Visualisation de la distribution

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

Diplôme le plus élevé du Chef de Ménage	**N = 7,156**[1]
Diplôme 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 (%)

Activité des 7 derniers jours du Chef de Ménage	**N = 7,156**[1]
Activité 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 (%)

```
# Affichage des labels de la variable hactiv7j en 2021
labelled::val_labels(welf2021$hactiv7j)
```

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

```
# Résumé statistique de la variable hactiv7j dans la base 2018
welf2018 %>%
  to_factor() %>%
  select(hactiv7j) %>%
  tbl_summary(missing = "always", missing_text = "NA",label = list(hactiv7j ~ "Activité du CM"))%>%
    modify_header(label = "**Activité des 7 derniers jours du Chef de Ménage**")  # Génération du table
```

Activité des 7 derniers jours du Chef de Ménage	**N = 7,156**[1]
Activité 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 (%)

Recherche de l'incoherence Il y a une incohérence dans l'ordre des modalités entre les bases 2018 et 2021, bien que les labels soient identiques. Cela suggère une erreur dans le codage des valeurs. Une harmonisation des modalités est nécessaire pour assurer la comparabilité des données entre les deux années.

Correction de l'incohérence Pour corriger cette incohérence, il est nécessaire de réorganiser les modalités afin qu'elles correspondent exactement entre les deux bases. Une fois l'ordre des modalités ajusté dans welfare 2021, il sera alors possible d'affecter ses labels à welfare 2018 pour assurer une harmonisation complète.

```
welf2018 <- welf2018 %>%
  mutate(hactiv7j = dplyr::recode(hactiv7j,
                           `2` = 4,
                           `3` = 2,
                           `4` = 3))
```

```
# Affectation des labels de la variable 'hactiv7j' de la base 2021 à la base 2018
val_labels(welf2018$hactiv7j) <-  val_labels(welf2021$hactiv7j)
```

```
welf2018 %>%
  to_factor() %>%
  select(hactiv7j) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(hactiv7j ~ "Activité du CM"))%>%
              modify_header(label = "**Activité des 7 derniers jours du Chef de Ménage**")
```

Controle des modifications

d) Traitement de la variable hbranch

```
# Visualisation des labels de la variable 'hbranch' dans la base 2018
labelled::val_labels(welf2018$hbranch)
```

Branche d'activité du Chef de Ménage	**N = 7,156**[1]
Branche activité du CM	
Agriculture	1,366 (25%)
Elevage/peche	374 (6.9%)
Indust. extr.	58 (1.1%)
Autr. indust.	497 (9.1%)
btp	313 (5.8%)
Commerce	1,094 (20%)
Restaurant/Hotel	63 (1.2%)
Trans./Comm.	251 (4.6%)
Education/Sante	379 (7.0%)
Services perso.	761 (14%)
Aut. services	278 (5.1%)
NA	1,722

[1] n (%)

Visualisation de la distribution

```
##      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
```

```
# Visualisation des labels de la variable 'hbranch' dans la base 2021
labelled::val_labels(welf2021$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
```

```
# Vérification de la distribution des valeurs manquantes pour la variable 'hbranch' dans la base 2018
welf2018 %>%
  to_factor() %>%
  select(hbranch) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(hbranch ~ "Branche activité du CM"))%>%
              modify_header(label = "**Branche d'activité du Chef de Ménage**")
```

Recherche de l'incoherence Il est observé qu'en 2021, une nouvelle modalité correspondant au secteur de la sylvopasture a été ajoutée et combinée avec les secteurs de l'élevage et de la pêche, ce qui diffère de la catégorisation de 2018.

Branche d'activité du Chef de Ménage	**N = 7,156**[1]
Branche activité du CM	
Agriculture	1,366 (25%)
Elevage/syl./peche	374 (6.9%)
Indust. extr.	58 (1.1%)
Autr. indust.	497 (9.1%)
btp	313 (5.8%)
Commerce	1,094 (20%)
Restaurant/Hotel	63 (1.2%)
Trans./Comm.	251 (4.6%)
Education/Sante	379 (7.0%)
Services perso.	761 (14%)
Aut. services	278 (5.1%)
NA	1,722

[1] n (%)

Correction de l'incohérence Cette incohérence peut être corrigée par une simple affectation, similaire à celle effectuée pour les autres variables.

```
# Affectation des labels de la variable 'hactiv7j' de la base 2021 à la base 2018
val_labels(welf2018$hbranch) <-  val_labels(welf2021$hbranch)
```

```
welf2018 %>%
  to_factor() %>%
  select(hbranch) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(hbranch ~ "Branche activité du CM"))%>%
              modify_header(label = "**Branche d'activité du Chef de Ménage**")
```

Controle des modifications

v) Traitement de la variable hcsp

```
# Affichage des labels de la variable hcsp pour la base welfare 2018
labelled::val_labels(welf2018$hcsp)
```

Visualisation de la distribution

```
##                                          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
```

```
# Affichage des labels de la variable hcsp pour la base welfare 2021
labelled::val_labels(welf2021$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
```

Recherche de l'incoherence Comme la différence réside uniquement dans la labellisation de la neuvième modalité,il suffit de mettre à jour les labels de la base welfare 2021 pour les aligner sur ceux de 2018.

Correction de l'incohérence Ainsi, les labels de 2018 seront affectés à ceux de 2021.

```
#Affectation des labels de 2018 à 2021
val_labels(welf2021$hcsp) <-  val_labels(welf2018$hcsp)
```

Catégorie Socioprofessionnelle du Chef de Ménage	**N = 7,120**[1]
Catégorie SP du CM	
Cadre supérieur	57 (1.0%)
Cadre moyen/agent de maîtrise	280 (4.8%)
Ouvrier ou employé qualifié	450 (7.8%)
Ouvrier ou employé non qualifié	332 (5.7%)
Manœuvre, aide ménagère	151 (2.6%)
Stagiaire ou Apprenti rénuméré	34 (0.6%)
Stagiaire ou Apprenti non rénuméré	3 (<0.1%)
Travailleur familial contribuant à une entreprise familiale	66 (1.1%)
Travailleur pour compte propre	4,302 (74%)
Patron	119 (2.1%)
NA	1,326

[1] n (%)

```
welf2021 %>%
  to_factor() %>%
  select(hcsp) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(hcsp ~ "Catégorie SP du CM"))%>%
              modify_header(label = "**Catégorie Socioprofessionnelle du Chef de Ménage**")
```

Controle des modifications

e) Traitement de la variable zae

```
# Vérification des labels de la variable zae dans la base welfare 2018
labelled::val_labels(welf2018$zae)
```

Visualisation de la distribution

```
## NULL
```

```
# Vérification des labels de la variable zae dans la base welfare 2021
labelled::val_labels(welf2021$zae)
```

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

Zone Agroécologique	**N = 7,156**[1]
ZAE	
1	1,020 (14%)
2	912 (13%)
3	1,602 (22%)
4	1,414 (20%)
5	1,752 (24%)
6	456 (6.4%)
NA	0

[1] n (%)

```
# Résumé de la distribution de la variable zae dans la base welfare 2018, avec gestion des valeurs manq
welf2018 %>%
  to_factor() %>%
  select(zae) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(zae ~ "ZAE"))%>%
              modify_header(label = "**Zone Agroécologique**")
```

```
# Résumé de la distribution de la variable zae dans la base welfare 2021, avec gestion des valeurs manq
welf2021 %>%
  select(zae) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(zae ~ "ZAE"))%>%
              modify_header(label = "**Zone Agroécologique**")
```

```
## ! Column(s) "zae" are class "haven_labelled".
## i This is an intermediate data structure not meant for analysis.
## i Convert columns with `haven::as_factor()`, `labelled::to_factor()`,
##   `labelled::unlabelled()`, and `unclass()`. Failure to convert may have
##   unintended consequences or result in error.
## <https://haven.tidyverse.org/articles/semantics.html>
## <https://larmarange.github.io/labelled/articles/intro_labelled.html#unlabelled>
```

Identification de l'incohérence Il a été observé que dans la base welfare 2018, la variable zae n'est pas labellisée, et les codes diffèrent entre les deux bases. En 2018, les codes sont 1, 2, 3, 4, 5 et 6, tandis qu'en 2021, ils sont 1, 3, 5, 7, 9 et 11. De plus, les codes correspondant à Kedougou et Dakar sont inversés dans la base 2018, bien qu'il faille prendre en compte les fréquences de chaque code dans les deux bases.

Correction de l'incohérence Il sera d'abord nécessaire d'aligner les codes de 2018 avec ceux de 2021. Une fois cette correspondance effectuée, il sera possible d'affecter les labels appropriés à la variable dans la base welfare 2018, en suivant le même processus que pour la base welfare 2021.

Zone Agroécologique	**N = 7,120**[1]
ZAE	
1	452 (6.3%)
3	911 (13%)
5	1,599 (22%)
7	1,413 (20%)
9	1,740 (24%)
11	1,005 (14%)
NA	0

[1] n (%)

Rétablissement de l'ordre pour Kedougou et Dakar en 2018 Afin de corriger l'inversion des codes pour Kedougou et Dakar dans la base 2018, il convient d'échanger les codes correspondants dans cette base. Cette modification permettra d'aligner l'ordre des modalités sur celui de la base 2021, garantissant ainsi la cohérence des données.“‘{r}

```
#Correction de l'inversion de codes pour Kedougou et Dakar en 2018
#Ce code permet d'inverser les codes pour Kedougou et Dakar dans la base de données 2018 afin de rétabl
welf2018 <- welf2018 %>%
  mutate(zae = dplyr::recode(zae,
                             '1' = 6, # Kédougou remplace Dakar
                             '2' = 2,
                             '3' = 3,
                             '4' = 4,
                             '5' = 5,
                             '6' = 1   # Dakar devient Kedougou
                             ))
```

```
#Vérification après correction des codes de Kedougou et Dakar en 2018
#Ce premier bloc permet de vérifier l'effet de la mutation dans la base de données 2018 après la correc
welf2018 %>%
  to_factor() %>%
  select(zae) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(zae ~ "ZAE"))%>%
              modify_header(label = "**Zone Agroécologique**")
```

```
#Harmonisation des codes géographiques dans la base 2021
#Le deuxième bloc effectue l'harmonisation des codes géographiques dans la base 2021 en réaffectant les
welf2021 <- welf2021 %>%
  mutate(zae = dplyr::recode(zae,
    `1` = 1,
    `3` = 2,
```

Zone Agroécologique	**N = 7,156**[1]
ZAE	
1	456 (6.4%)
2	912 (13%)
3	1,602 (22%)
4	1,414 (20%)
5	1,752 (24%)
6	1,020 (14%)
NA	0

[1] n (%)

```
    `5` = 3,
    `7` = 4,
    `9` = 5,
    `11` = 6
  ))
```

Vérification après correction effectuée entre Kedougou et Dakar

```
#Vérification du recodage en 2021
welf2021 %>%
  select(zae) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(zae ~ "ZAE"))%>%
              modify_header(label = "**Zone Agroécologique**")
```

Controle des modifications

```
## ! Column(s) "zae" are class "haven_labelled".
## i This is an intermediate data structure not meant for analysis.
## i Convert columns with `haven::as_factor()`, `labelled::to_factor()`,
##   `labelled::unlabelled()`, and `unclass()`. Failure to convert may have
##   unintended consequences or result in error.
## <https://haven.tidyverse.org/articles/semantics.html>
## <https://larmarange.github.io/labelled/articles/intro_labelled.html#unlabelled>
```

```
# Recodage des modalités de `zae` dans la base de données 2021
welf2021 <- welf2021 %>%
  mutate(zae = dplyr::recode(zae,
    `1` = "Kédougou",
    `2` = "Saint-Louis-Matam",
    `3` = "Thies-Diourbel-Louga",
    `4` = "Kaolack-Fatick-Kaffrine",
    `5` = "Ziguinchor-Tamba-Kolda-Sédhiou",
```

Zone Agroécologique	**N = 7,120**[1]
ZAE	
1	452 (6.3%)
2	911 (13%)
3	1,599 (22%)
4	1,413 (20%)
5	1,740 (24%)
6	1,005 (14%)
NA	0

[1] n (%)

```
    `6` = "Dakar"
  ))
```

```
# Recodage des modalités de `zae` dans la base de données 2018
welf2018 <- welf2018 %>%
  mutate(zae = dplyr::recode(zae,
    `1` = "Kédougou",
    `2` = "Saint-Louis-Matam",
    `3` = "Thies-Diourbel-Louga",
    `4` = "Kaolack-Fatick-Kaffrine",
    `5` = "Ziguinchor-Tamba-Kolda-Sédhiou",
    `6` = "Dakar"
  ))
```

```
# Vérification des changements dans la base de données 2018
welf2018 %>%
  to_factor() %>%
  select(zae) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(zae ~ "ZAE"))%>%
              modify_header(label = "**Zone Agroécologique**")
```

```
# Vérification des changements dans la base de données 2021
welf2021 %>%
  to_factor() %>%
  select(zae) %>%
  tbl_summary(missing = "always",
              missing_text = "NA",
              label = list(zae ~ "ZAE"))%>%
              modify_header(label = "**Zone Agroécologique**")
```

Zone Agroécologique	**N = 7,156**[1]
ZAE	
Dakar	1,020 (14%)
Kaolack-Fatick-Kaffrine	1,414 (20%)
Kédougou	456 (6.4%)
Saint-Louis-Matam	912 (13%)
Thies-Diourbel-Louga	1,602 (22%)
Ziguinchor-Tamba-Kolda-Sédhiou	1,752 (24%)
NA	0

[1] n (%)

Zone Agroécologique	**N = 7,120**[1]
ZAE	
Dakar	1,005 (14%)
Kaolack-Fatick-Kaffrine	1,413 (20%)
Kédougou	452 (6.3%)
Saint-Louis-Matam	911 (13%)
Thies-Diourbel-Louga	1,599 (22%)
Ziguinchor-Tamba-Kolda-Sédhiou	1,740 (24%)
NA	0

[1] n (%)

Vérification des modifications Le codage et la labellisation de la variable zae sont désormais cohérents entre les deux bases. Il reste à vérifier qu'aucune différence n'existe dans la labellisation ou le codage des autres variables communes aux deux bases.

3.Controle du traitement des données

```
discord_lab_var <- c()

for (var in var_communes) { # Parcourir les variables communes entre les deux bases

  # Vérifier si la variable est labellisée dans la base 2018
  if (labelled::is.labelled(welf2018[[var]])) {
    lab_val2018 <- labelled::val_labels(welf2018[[var]]) # Récupérer les labels de la variable en 2018
  } else {
    lab_val2018 <- NULL # Assigner NULL si la variable n'est pas labellisée en 2018
  }

  # Vérifier si la variable est labellisée dans la base 2021
  if (labelled::is.labelled(welf2021[[var]])) {
    lab_val2021 <- labelled::val_labels(welf2021[[var]]) # Récupérer les labels de la variable en 2021
  } else {
    label_val_2021 <- NULL # Assigner NULL si la variable n'est pas labellisée en 2021
```

```
  }

  # Comparer les labels des deux bases
  if (!identical(lab_val2018, lab_val2021)) {
    discord_lab_var <- append(discord_lab_var, var)
    print(var)
  }
}
```

```
## [1] "country"
## [1] "year"
## [1] "hhid"
## [1] "grappe"
## [1] "menage"
## [1] "vague"
## [1] "zae"
## [1] "hhweight"
## [1] "hhsize"
## [1] "eqadu1"
## [1] "eqadu2"
## [1] "hage"
## [1] "dali"
## [1] "dnal"
## [1] "dtot"
## [1] "pcexp"
## [1] "zzae"
## [1] "zref"
## [1] "def_spa"
## [1] "def_temp"
```

Fusion des bases welfare 2018 et 2021

Après avoir harmonisé le codage et la labellisation des variables communes entre les deux bases, la prochaine étape consiste à les fusionner. Cette fusion permettra de créer une base de données consolidée, comprenant les informations des deux années, afin d'effectuer des analyses comparatives et des traitements ultérieurs sur les données combinées.

```
welf_finale <- bind_rows(welf2018, welf2021)
```

Récupération de la nouvelle base de données

```
# Exporter la base de données finale en dta
write_dta(welf_finale, "../Sorties/welf_finale.dta")
```