TP9 : Fusion et harmonisation des bases de données welfare de l’EHCVM 2018 et 2021 au Sénégal

**RASAMOELINA Paulinah**

2025-04-02

# **Section 0 : Installation, Importation des packages**

Dans cette section, nous installons et importons tous les packages nécessaires.

packages <- c("readr","haven","utils","dplyr","labelled","questionr","gtsummary","ggplot2", "lmtest", "sandwich", "plm")  
  
  
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 ()  
}

# **Section 1 : Comprendre le problème**

## **Objectif du travail**

L’objectif de ce travail est de fusionner deux bases de données distinctes, à savoir **ehcvm\_welfare\_sen2018** et **ehcvm\_welfare\_sen2021**, afin de créer une base de données consolidée intitulée **ehcvm\_welfare\_sen2018\_2021**, qui sera sauvegardée au format **.DTA**. Cette nouvelle base de données combinée permettra d’analyser les tendances du bien-être au Sénégal en 2018 et 2021, en facilitant les comparaisons et les analyses longitudinales.

## **Variables clés**

Les variables clés de cette analyse sont principalement celles qui nous permettront de mesurer les caractéristiques des ménages et des individus. Parmi les variables importantes, nous pouvons inclure :

Taille du ménage (hhsize) : Elle représente le nombre de personnes vivant dans chaque ménage, ce qui est crucial pour comprendre la structure démographique et les conditions de vie.

Âge du chef de ménage (hage) : Cette variable permet d’analyser l’âge des chefs de ménage et son impact sur les caractéristiques socio-économiques du ménage.

Dépenses par habitant (pcexp) : Cela mesure les dépenses de chaque ménage par habitant et est essentiel pour évaluer le niveau de vie.

Revenus ou sources de revenus (hactiv7j) : Identifie les principales sources de revenus des ménages, ce qui permet d’analyser la pauvreté et les inégalités.

Niveau d’éducation du chef de ménage (heduc) : Cela permet d’analyser l’impact de l’éducation sur le bien-être des ménages.

## **Contraintes spécifiques**

Problèmes d’harmonisation des données : Les deux bases de données proviennent de différentes années (2018 et 2021), ce qui peut entraîner des différences dans la structure des variables. Il est important de vérifier et d’harmoniser les variables avant de les combiner pour assurer une analyse cohérente.

Valeurs manquantes : Il peut y avoir des valeurs manquantes dans les deux bases de données. Il faudra décider comment les traiter, soit en les imitant, soit en les excluant selon leur impact sur l’analyse.

Consistance des unités et des définitions : Il est essentiel de s’assurer que les unités de mesure (par exemple, les dépenses) et les définitions des variables sont cohérentes entre les deux années.

Respect de la confidentialité et des données sensibles : Assurez-vous que l’utilisation des données respecte les règles de confidentialité et les lois sur la protection des données personnelles.

# **Section 2 : Chargement des données**

## **Lecture des fichiers au format .DTA**

# La base welfare\_sen2018.  
  
welfare\_sen2018 <- haven::read\_dta("../donnees/ehcvm\_welfare\_sen2018.dta")  
  
# La base welfare\_sen2021.  
welfare\_sen2021 <- haven::read\_dta("../donnees/ehcvm\_welfare\_sen2021.dta")

## **Aperçu des premières lignes de deux bases de donnéess**

# Affichage des premières lignes des bases welfare\_sen2018 et welfare\_sen2021.  
  
print("Les premières lignes de la base welfare\_sen2018")

## [1] "Les premières lignes de la base welfare\_sen2018"

head(welfare\_sen2018)

## # A tibble: 6 × 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>

print("Les premières lignes de la base welfare\_sen2021")

## [1] "Les premières lignes de la base welfare\_sen2021"

head(welfare\_sen2021)

## # A tibble: 6 × 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 de la dimension de deux bases**

# Nombre de variables dans welfare\_sen2018  
print("Nombre de variables dans welfare\_sen2018 :")

## [1] "Nombre de variables dans welfare\_sen2018 :"

ncol(welfare\_sen2018)

## [1] 35

# Nombre de variables dans welfare\_sen2021  
print("Nombre de variables dans welfare\_sen2021 :")

## [1] "Nombre de variables dans welfare\_sen2021 :"

ncol(welfare\_sen2021)

## [1] 47

## **Informations générales sur les colonnes/variables**

# Liste des noms des colonnes des bases welfare\_sen2018 et welfare\_sen2021.  
  
print("Liste des noms des colonnes de la base welfare\_sen2018")

## [1] "Liste des noms des colonnes de la base welfare\_sen2018"

colnames(welfare\_sen2018)

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

print("Liste des noms des colonnes de la base welfare\_sen2021")

## [1] "Liste des noms des colonnes de la base welfare\_sen2021"

colnames(welfare\_sen2021)

## [1] "grappe" "menage" "country"   
## [4] "year" "hhid" "vague"   
## [7] "month" "zae" "region"   
## [10] "milieu" "hhweight" "hhsize"   
## [13] "eqadu1" "eqadu2" "hgender"   
## [16] "hage" "hmstat" "hreligion"   
## [19] "hnation" "hethnie" "halfa"   
## [22] "halfa2" "heduc" "hdiploma"   
## [25] "hhandig" "hactiv7j" "hactiv12m"   
## [28] "hbranch" "hsectins" "hcsp"   
## [31] "dali" "dnal" "dtot"   
## [34] "pcexp" "zzae" "zref"   
## [37] "def\_spa" "def\_temp" "def\_temp\_prix2021m11"  
## [40] "def\_temp\_cpi" "def\_temp\_adj" "zali0"   
## [43] "dtet" "monthly\_cpi" "cpi2017"   
## [46] "icp2017" "dollars"

# **Section 3 : Verification de la qualité des données**

Dans cette section, nous procéderons à la verification, à l’analyse et à la comparaison des deux bases afin d’identifier leurs similitudes et leurs différences, ce qui nous permettra de mieux les exploiter par la suite.

## **Valeurs manquantes dans les bases welfare\_sen2018 et welfare\_sen2021**

# Calcul du nombre de valeurs manquantes par variable (colonne) dans les bases welfare\_sen2018 et welfare\_sen2021  
  
print("Valeurs manquantes dans la base welfare\_sen2018 :")

## [1] "Valeurs manquantes dans la base welfare\_sen2018 :"

welfare\_missing\_sen2018 <- colSums(is.na(welfare\_sen2018))  
welfare\_missing\_sen2018

## 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 dans la base welfare\_sen2021 :")

## [1] "Valeurs manquante dans la base welfare\_sen2021 :"

welfare\_missing\_sen2021 <- colSums(is.na(welfare\_sen2021))  
welfare\_missing\_sen2021

## 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, il n’y a pas beaucoup de valeurs manquantes. Seules quatre variables présentent des valeurs manquantes en 2021 et quatre variables en 2018.

## **Vérification des doublons**

Nous allons vérifier la présence de doublons dans les bases welfare\_sen2018 et welfare\_sen2021 afin de garantir la qualité des données avant toute analyse.

duplicate\_2018 <- welfare\_sen2018[duplicated(welfare\_sen2018[, c("grappe","menage")]), ] # Sélection des ménages dupliqués dans la base welfare\_sen2018  
  
duplicate\_2021 <- welfare\_sen2021[duplicated(welfare\_sen2021[, c("grappe","menage")]), ] # Sélection des ménages dupliqués dans la base welfare\_sen2021  
  
cat("Nombre de doublons présents dans la base welfare\_sen2018 :", nrow(duplicate\_2018), "\n")

## Nombre de doublons présents dans la base welfare\_sen2018 : 0

cat("Nombre de doublons présents dans la base welfare\_sen2021 :", nrow(duplicate\_2021))

## Nombre de doublons présents dans la base welfare\_sen2021 : 0

Aucune des deux bases ne contient de ménages dupliqués.

## **Types de données dans les deux bases de données**

# Structure des variables des bases welfare\_sen2018 et welfare\_sen2021.  
  
print("La structure des variables de la base welfare\_sen2018")

## [1] "La structure des variables de la base welfare\_sen2018"

utils::str(welfare\_sen2018)

## tibble [7,156 × 35] (S3: tbl\_df/tbl/data.frame)  
## $ country : chr [1:7156] "SEN" "SEN" "SEN" "SEN" ...  
## ..- attr(\*, "label")= chr "Pays"  
## ..- attr(\*, "format.stata")= chr "%3s"  
## $ year : num [1:7156] 2018 2018 2018 2018 2018 ...  
## ..- attr(\*, "label")= chr "Annee enquete"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ hhid : num [1:7156] 1001 1002 1003 2001 2002 ...  
## ..- attr(\*, "label")= chr "Idenfiant menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ grappe : num [1:7156] 1 1 1 2 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "Numero grappe"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ menage : num [1:7156] 1 2 3 1 2 3 4 5 6 7 ...  
## ..- attr(\*, "label")= chr "Numero menage"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ vague : num [1:7156] 1 1 1 2 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "Vague"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ zae : num [1:7156] 1 1 1 1 1 1 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "Zone agroecologique"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ region : dbl+lbl [1:7156] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Region residence"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:14] 1 2 3 4 5 6 7 8 9 10 ...  
## .. ..- attr(\*, "names")= chr [1:14] "dakar" "ziguinchor" "diourbel" "SAINT-LOUIS" ...  
## $ milieu : dbl+lbl [1:7156] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Milieu residence"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 1 2  
## .. ..- attr(\*, "names")= chr [1:2] "Urbain" "Rural"  
## $ hhweight : num [1:7156] 1750 1750 1750 266 266 ...  
## ..- attr(\*, "label")= chr "Ponderation menage"  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ hhsize : num [1:7156] 2 2 1 10 6 4 3 2 3 1 ...  
## ..- attr(\*, "label")= chr "Taille menage"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ eqadu1 : num [1:7156] 1.32 1.45 1 7.66 4.28 ...  
## ..- attr(\*, "label")= chr "Nbr adultes-equiv. FAO"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ eqadu2 : num [1:7156] 1.61 1.61 1 5.54 3.25 ...  
## ..- attr(\*, "label")= chr "Nbr adultes-equiv. alt."  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ hgender : dbl+lbl [1:7156] 2, 1, 1, 1, 1, 2, 1, 2, 2, 2, 1, 1, 2, 1, 2, 1, 1, 1,...  
## ..@ label : chr "Genre du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 1 2  
## .. ..- attr(\*, "names")= chr [1:2] "Masculin" "Féminin"  
## $ hage : num [1:7156] 59 59 27 85 40 68 31 54 67 42 ...  
## ..- attr(\*, "label")= chr "Age du CM"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ hmstat : dbl+lbl [1:7156] 1, 2, 1, 2, 2, 5, 1, 2, 5, 1, 2, 2, 2, 1, 2, 2, 2, 6,...  
## ..@ label : chr "Situation famille du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:7] 1 2 3 4 5 6 7  
## .. ..- attr(\*, "names")= chr [1:7] "Célibataire" "Marié(e) monogame" "Marié(e) polygame" "Union libre" ...  
## $ hreligion: dbl+lbl [1:7156] 1, 2, 2, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Religion du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:5] 1 2 3 4 5  
## .. ..- attr(\*, "names")= chr [1:5] "Musulman" "Chrétien" "Animiste" "Autre Réligion" ...  
## $ hnation : dbl+lbl [1:7156] 7, 12, 7, 7, 7, 7, 11, 7, 7, 7, 7, 7, 7, 1...  
## ..@ label : chr "Nationalite du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:12] 1 2 3 4 5 6 7 8 9 10 ...  
## .. ..- attr(\*, "names")= chr [1:12] "Benin" "Burkina Faso" "Côte d'Ivoire" "Guinée Bissau" ...  
## $ halfab : dbl+lbl [1:7156] 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0,...  
## ..@ label : chr "Alphabetisation du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 0 1  
## .. ..- attr(\*, "names")= chr [1:2] "Non" "Oui"  
## $ heduc : dbl+lbl [1:7156] 3, 9, 9, 3, 1, 1, 9, 1, 3, 4, 3, 1, 1, 1, 1, 9, 9, 1,...  
## ..@ label : chr "Education du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:9] 1 2 3 4 5 6 7 8 9  
## .. ..- attr(\*, "names")= chr [1:9] "Aucun" "Maternelle" "Primaire" "Second. gl 1" ...  
## $ hdiploma : dbl+lbl [1:7156] 1, 6, 10, 0, 0, 0, 10, 0, 0, 1, 0, 0, 0, ...  
## ..@ label : chr "Diplome du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:11] 0 1 2 3 4 5 6 7 8 9 ...  
## .. ..- attr(\*, "names")= chr [1:11] "Aucun" "CEP/CFEE" "BEPC/BFEM" "cap" ...  
## $ hhandig : dbl+lbl [1:7156] 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## ..@ label : chr "Handicap majeur CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 0 1  
## .. ..- attr(\*, "names")= chr [1:2] "Non" "Oui"  
## $ hactiv7j : dbl+lbl [1:7156] 1, 1, 1, 5, 1, 5, 1, 1, 5, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Activite 7 jours du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:6] 1 2 3 4 5 6  
## .. ..- attr(\*, "names")= chr [1:6] "Occupe" "Chomeur" "TF cherchant emploi" "TF cherchant pas" ...  
## $ hactiv12m: dbl+lbl [1:7156] 1, 1, 1, 3, 1, 3, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Activite 12 mois du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:4] 1 2 3 4  
## .. ..- attr(\*, "names")= chr [1:4] "Occupe" "Trav. fam." "Non occupe" "Moins de 5 ans"  
## $ hbranch : dbl+lbl [1:7156] 6, 6, 11, NA, 6, NA, 9, 11, NA, 6, 7, 8, 6, 1...  
## ..@ label : chr "Branche activite du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:11] 1 2 3 4 5 6 7 8 9 10 ...  
## .. ..- attr(\*, "names")= chr [1:11] "Agriculture" "Elevage/peche" "Indust. extr." "Autr. indust." ...  
## $ hsectins : dbl+lbl [1:7156] 3, 3, 1, NA, 3, NA, 1, 5, NA, 3, 3, 3, 3, ...  
## ..@ label : chr "Secteur instit. du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:6] 1 2 3 4 5 6  
## .. ..- attr(\*, "names")= chr [1:6] "Etat/Collectivités locales" "Entreprise publique/ parapublique" "Entreprise Privée" "Entreprise associative" ...  
## $ hcsp : dbl+lbl [1:7156] 4, 1, 7, NA, 4, NA, 6, 5, NA, 9, 3, 4, 9, ...  
## ..@ label : chr "CSP du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:10] 1 2 3 4 5 6 7 8 9 10  
## .. ..- attr(\*, "names")= chr [1:10] "Cadre supérieur" "Cadre moyen/agent de maîtrise" "Ouvrier ou employé qualifié" "Ouvrier ou employé non qualifié" ...  
## $ dali : num [1:7156] 1978488 1886313 387989 6568518 2362250 ...  
## ..- attr(\*, "label")= chr "Conso annuelle alim. menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ dnal : num [1:7156] 13139233 8894985 5055663 4889539 3930703 ...  
## ..- attr(\*, "label")= chr "Conso annuelle non alim. menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ dtot : num [1:7156] 15117721 10781298 5443652 11458057 6292954 ...  
## ..- attr(\*, "label")= chr "Conso annuelle totale menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ pcexp : num [1:7156] 6440507 4593088 4638249 976281 893649 ...  
## ..- attr(\*, "label")= chr "Indicateur de bien-être"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ zzae : num [1:7156] 391340 391340 391340 391340 391340 ...  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ zref : num [1:7156] 333441 333441 333441 333441 333441 ...  
## ..- attr(\*, "label")= chr "Seuil pauvrete national"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ def\_spa : num [1:7156] 1.17 1.17 1.17 1.17 1.17 ...  
## ..- attr(\*, "label")= chr "Deflateur spatial"  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ def\_temp : num [1:7156] 1.015 1.012 1.015 0.992 0.992 ...  
## ..- attr(\*, "label")= chr "Deflateur temporel"  
## ..- attr(\*, "format.stata")= chr "%9.0g"

print("La structure des variables de la base welfare\_sen2021")

## [1] "La structure des variables de la base welfare\_sen2021"

utils::str(welfare\_sen2021)

## tibble [7,120 × 47] (S3: tbl\_df/tbl/data.frame)  
## $ grappe : num [1:7120] 2 2 2 2 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "grappe"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ menage : num [1:7120] 5 15 3 13 8 16 7 4 1 12 ...  
## ..- attr(\*, "label")= chr "Identifiant du ménage"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ country : chr [1:7120] "SEN" "SEN" "SEN" "SEN" ...  
## ..- attr(\*, "label")= chr "Pays"  
## ..- attr(\*, "format.stata")= chr "%3s"  
## $ year : num [1:7120] 2021 2021 2021 2021 2021 ...  
## ..- attr(\*, "label")= chr "Annee enquete"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ hhid : num [1:7120] 205 215 203 213 208 216 207 204 201 212 ...  
## ..- attr(\*, "label")= chr "Idenfiant menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ vague : num [1:7120] 2 2 2 2 2 2 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "Vague"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ month : Date[1:7120], format: "2022-05-01" "2022-05-01" ...  
## $ zae : dbl+lbl [1:7120] 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 1...  
## ..@ label : chr "Zone agroecologique"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:6] 1 3 5 7 9 11  
## .. ..- attr(\*, "names")= chr [1:6] "Kédougou" "Saint-Louis-Matam" "Thies-Diourbel-Louga" "Kaolack-Fatick-Kaffrine" ...  
## $ region : dbl+lbl [1:7120] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Region residence"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:14] 1 2 3 4 5 6 7 8 9 10 ...  
## .. ..- attr(\*, "names")= chr [1:14] "dakar" "ziguinchor" "diourbel" "SAINT-LOUIS" ...  
## $ milieu : dbl+lbl [1:7120] 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Milieu residence"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 1 2  
## .. ..- attr(\*, "names")= chr [1:2] "Urbain" "Rural"  
## $ hhweight : num [1:7120] 386 386 386 386 290 ...  
## ..- attr(\*, "label")= chr "Ponderation menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ hhsize : num [1:7120] 3 2 3 5 1 5 1 3 9 6 ...  
## ..- attr(\*, "label")= chr "Taille menage"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ eqadu1 : num [1:7120] 2.28 1.66 2.82 4 0.79 ...  
## ..- attr(\*, "label")= chr "Nbr adultes-equiv. FAO"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ eqadu2 : num [1:7120] 2.2 1.61 1.87 3.17 1 ...  
## ..- attr(\*, "label")= chr "Nbr adultes-equiv. alt."  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ hgender : dbl+lbl [1:7120] 2, 1, 2, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1,...  
## ..@ label : chr "Genre du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 1 2  
## .. ..- attr(\*, "names")= chr [1:2] "Masculin" "Féminin"  
## $ hage : num [1:7120] 33 42 47 56 54 45 46 34 88 51 ...  
## ..- attr(\*, "label")= chr "Age du CM"  
## ..- attr(\*, "format.stata")= chr "%8.0g"  
## $ hmstat : dbl+lbl [1:7120] 3, 1, 3, 2, 6, 2, 1, 1, 2, 2, 3, 5, 1, 5, 3, 2, 2, 1,...  
## ..@ label : chr "Situation famille du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:7] 1 2 3 4 5 6 7  
## .. ..- attr(\*, "names")= chr [1:7] "Célibataire" "Marié(e) monogame" "Marié(e) polygame" "Union libre" ...  
## $ hreligion : dbl+lbl [1:7120] 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Religion du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:5] 1 2 3 4 5  
## .. ..- attr(\*, "names")= chr [1:5] "Musulman" "Chrétien" "Animiste" "Autre Réligion" ...  
## $ hnation : dbl+lbl [1:7120] 13, 13, 13, 13, 13, 13, 13, 17, 13, 13, 13, 13, 13, 1...  
## ..@ label : chr "Nationalite du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:17] 1 2 3 4 5 6 7 8 9 10 ...  
## .. ..- attr(\*, "names")= chr [1:17] "Bénin" "Burkina Faso" "Cape-vert" "Cote d'ivoire" ...  
## $ hethnie : dbl+lbl [1:7120] 2, 3, 6, 1, 1, 1, 5, NA, 8, 2, 1, 1, 1, ...  
## ..@ label : chr "Ethnie du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:13] 1 2 3 4 5 6 7 8 9 10 ...  
## .. ..- attr(\*, "names")= chr [1:13] "Wolof/Lébou" "Sérère" "Poular" "Soninké" ...  
## $ halfa : dbl+lbl [1:7120] 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1,...  
## ..@ label : chr "Alpha. lire/ecr. CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 0 1  
## .. ..- attr(\*, "names")= chr [1:2] "Non" "Oui"  
## $ halfa2 : dbl+lbl [1:7120] 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1,...  
## ..@ label : chr "Alpha. lire/ecr./comp. CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 0 1  
## .. ..- attr(\*, "names")= chr [1:2] "Non" "Oui"  
## $ heduc : dbl+lbl [1:7120] 1, 9, 4, 3, 1, 9, 4, 9, 3, 1, 1, 3, 4, 3, 1, 4, 9, 9,...  
## ..@ label : chr "Education du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:9] 1 2 3 4 5 6 7 8 9  
## .. ..- attr(\*, "names")= chr [1:9] "Aucun" "Maternelle" "Primaire" "Second. gl 1" ...  
## $ hdiploma : dbl+lbl [1:7120] 0, 7, 1, 1, 0, 6, 1, 10, 0, 0, 0, 0, 1, ...  
## ..@ label : chr "Diplome du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:11] 0 1 2 3 4 5 6 7 8 9 ...  
## .. ..- attr(\*, "names")= chr [1:11] "Aucun" "cepe" "bepc" "cap" ...  
## $ hhandig : dbl+lbl [1:7120] 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,...  
## ..@ label : chr "Handicap majeur CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:2] 0 1  
## .. ..- attr(\*, "names")= chr [1:2] "Non" "Oui"  
## $ hactiv7j : dbl+lbl [1:7120] 1, 1, 1, 1, 1, 1, 1, 5, 1, 1, 1, 5, 5, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Activite 7 jours du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:6] 1 2 3 4 5 6  
## .. ..- attr(\*, "names")= chr [1:6] "Occupe" "TF cherchant emploi" "TF cherchant pas" "Chomeur" ...  
## $ hactiv12m : dbl+lbl [1:7120] 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, 3, 1, 1, 1, 1, 1,...  
## ..@ label : chr "Activite 12 mois du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:4] 1 2 3 4  
## .. ..- attr(\*, "names")= chr [1:4] "Occupe" "Trav. fam." "Non occupe" "Moins de 5 ans"  
## $ hbranch : dbl+lbl [1:7120] 11, 9, 11, 11, 11, 9, 6, NA, 2, 4, 6, NA, NA, 1...  
## ..@ label : chr "Branche activite du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:11] 1 2 3 4 5 6 7 8 9 10 ...  
## .. ..- attr(\*, "names")= chr [1:11] "Agriculture" "Elevage/syl./peche" "Indust. extr." "Autr. indust." ...  
## $ hsectins : dbl+lbl [1:7120] 5, 3, 5, 3, 3, 2, 3, NA, 3, 3, 3, NA, NA, ...  
## ..@ label : chr "Secteur instit. du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:6] 1 2 3 4 5 6  
## .. ..- attr(\*, "names")= chr [1:6] "Etat/Collectivités locales" "Entreprise publique/ parapublique" "Entreprise Privée" "Entreprise associative" ...  
## $ hcsp : dbl+lbl [1:7120] 5, 3, 5, 4, 3, 3, 9, NA, 9, 9, 9, NA, NA, ...  
## ..@ label : chr "CSP du CM"  
## ..@ format.stata: chr "%8.0g"  
## ..@ labels : Named num [1:10] 1 2 3 4 5 6 7 8 9 10  
## .. ..- attr(\*, "names")= chr [1:10] "Cadre supérieur" "Cadre moyen/agent de maîtrise" "Ouvrier ou employé qualifié" "Ouvrier ou employé non qualifié" ...  
## $ dali : num [1:7120] 1357897 1330889 1510518 2759037 973303 ...  
## ..- attr(\*, "label")= chr "Conso annuelle alim. menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ dnal : num [1:7120] 930976 476051 1887723 2087600 1573509 ...  
## ..- attr(\*, "label")= chr "Conso annuelle non alim. menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ dtot : num [1:7120] 2288874 1806940 3398241 4846637 2546811 ...  
## ..- attr(\*, "label")= chr "Conso annuelle totale menage"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ pcexp : num [1:7120] 664906 787360 987171 844753 2219506 ...  
## ..- attr(\*, "label")= chr "Indicateur de bien-être"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ zzae : num [1:7120] 424179 424179 424179 424179 424179 ...  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ zref : num [1:7120] 369666 369666 369666 369666 369666 ...  
## ..- attr(\*, "label")= chr "Seuil pauvrete national"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ def\_spa : num [1:7120] 1.15 1.15 1.15 1.15 1.15 ...  
## ..- attr(\*, "label")= chr "Deflateur spatial"  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ def\_temp : num [1:7120] 0.991 0.991 0.991 0.991 1.013 ...  
## ..- attr(\*, "label")= chr "Deflateur temporel"  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ def\_temp\_prix2021m11: num [1:7120] 1.03 1.03 1.03 1.03 1.05 ...  
## ..- attr(\*, "label")= chr "temporal deflator for international poverty, 1 = 2021m11 prices"  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ def\_temp\_cpi : num [1:7120] 1 1 1 1 1.02 ...  
## ..- attr(\*, "label")= chr "alternative temporal deflator based on official CPI, 2018/19 style"  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ def\_temp\_adj : num [1:7120] 1.01 1.01 1.01 1.01 1.03 ...  
## ..- attr(\*, "label")= chr "temporal deflator adjusted for difference between hh and market survey periods"  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ zali0 : num [1:7120] 196233 196233 196233 196233 196233 ...  
## ..- attr(\*, "label")= chr "(sum) conso\_pc\_val\_up"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ dtet : num [1:7120] 769710 911466 1142771 977906 2515290 ...  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ monthly\_cpi : num [1:7120] 120 120 120 120 123 ...  
## ..- attr(\*, "label")= chr "Monthly CPI value"  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ cpi2017 : num [1:7120] 1.1 1.1 1.1 1.1 1.1 ...  
## ..- attr(\*, "format.stata")= chr "%9.0g"  
## $ icp2017 : num [1:7120] 239 239 239 239 239 ...  
## ..- attr(\*, "format.stata")= chr "%12.0g"  
## $ dollars : num [1:7120] 7.71 9.12 11.44 9.79 25.79 ...  
## ..- attr(\*, "label")= chr "welfare in 2017 PPP USD per capita per day (not spatially deflated)"  
## ..- attr(\*, "format.stata")= chr "%9.0g"

Par suite, nous cherchons à obtenir un aperçu complet de l’ensemble des variables issues des deux bases. Pour chaque variable, nous préciserons si elle est présente dans une seule base ou dans les deux bases.

vars\_2018\_base = colnames(welfare\_sen2018) # Liste des variables dans la base welfare\_sen2018.  
  
vars\_2021\_base = colnames(welfare\_sen2021) # Liste des variables dans la base welfare\_sen2021.  
  
# Identifions les variables communes entre les bases welfare\_sen2018 et welfare\_sen2021.  
  
common\_vars\_2018\_2021 <- intersect(vars\_2018\_base, vars\_2021\_base) # Prenons les variables qui sont présentes à la fois dans les bases welfare\_sen2018 et welfare\_sen2021.  
  
# Variables présentes uniquement dans la base welfare\_sen2018.  
  
vars\_base2018\_only <- setdiff(vars\_2018\_base, vars\_2021\_base) # Garde les variables de vars\_2018\_base qui ne sont pas présentes dans vars\_2021\_base.  
print("Variables présentes uniquement dans la base welfare\_sen2018 :")

## [1] "Variables présentes uniquement dans la base welfare\_sen2018 :"

vars\_base2018\_only

## [1] "halfab"

# Variables présentes uniquement en 2021  
  
vars\_base2021\_only <- setdiff(vars\_2021\_base, vars\_2018\_base) # Garde les variables de vars\_2021\_base qui ne sont pas présentes dans vars\_2018\_base.  
print("Variables présentes uniquement dans la base welfare\_sen2021 :")

## [1] "Variables présentes uniquement dans la base welfare\_sen2021 :"

vars\_base2021\_only

## [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"

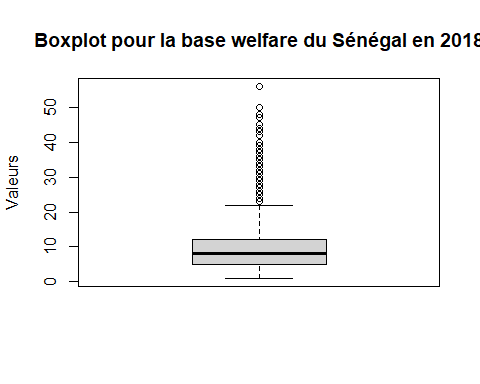
## **Detection de valeurs abérantes(Boxplot, IQR, Z-Score)**

Nous avons choisi la variable taille du ménage “hhsize” pour cette étude.

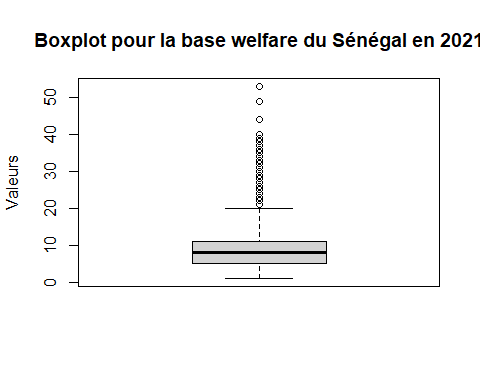
### **Boxplot**

Le Boxplot est un excellent moyen visuel de détecter les valeurs aberrantes. Il montre la distribution des données et les points qui se trouvent en dehors des “moustaches” sont considérés comme des outliers.

# Boxplot pour détecter les valeurs aberrantes dans une colonne de la base welfare\_sen2018  
boxplot(welfare\_sen2018$hhsize, main = "Boxplot pour la base welfare du Sénégal en 2018", ylab = "Valeurs")



# Boxplot pour welfare\_sen2021  
boxplot(welfare\_sen2021$hhsize, main = "Boxplot pour la base welfare du Sénégal en 2021", ylab = "Valeurs")



### **Méthode IQR (Intervalle Interquartile)**

La méthode IQR est utilisée pour identifier les outliers en fonction de l’intervalle interquartile, qui est la différence entre le troisième quartile (Q3) et le premier quartile (Q1). Les valeurs aberrantes sont généralement définies comme étant en dehors de l’intervalle : Q1 - 1,5*IQR et Q1 + 1,5*IQR

# Détection des valeurs aberrantes dans welfare\_sen2018 avec IQR  
Q1 <- quantile(welfare\_sen2018$hhsize, 0.25)  
Q3 <- quantile(welfare\_sen2018$hhsize, 0.75)  
IQR <- Q3 - Q1  
  
# Calcul des bornes  
borne\_inf <- Q1 - 1.5 \* IQR  
borne\_sup <- Q3 + 1.5 \* IQR  
  
# Identifier les valeurs aberrantes  
outliers <- welfare\_sen2018$hhsize[welfare\_sen2018$hhsize < borne\_inf | welfare\_sen2018$hhsize > borne\_sup]  
print(outliers)

## [1] 26 27 38 32 38 30 25 29 23 26 26 27 24 24 28 34 24 27 44 23 25 32 25 25 33  
## [26] 30 24 29 23 26 27 24 24 25 37 35 35 33 48 27 24 25 28 28 32 24 27 23 23 28  
## [51] 32 32 27 31 29 50 24 34 36 44 36 26 29 28 32 42 29 26 28 23 24 34 25 25 29  
## [76] 27 29 23 28 27 28 28 28 23 26 27 23 28 24 27 26 30 25 29 24 30 26 24 31 28  
## [101] 33 23 29 28 33 27 25 24 35 48 26 23 23 23 44 25 25 30 27 23 32 56 26 28 32  
## [126] 23 26 25 31 32 28 25 34 27 45 25 23 34 26 28 34 30 26 26 47 23 23 25 23 23  
## [151] 23 29 23 23 31 25 35 24 24 43 23 26 29 25 44 42 25 37 40 25 42 33 44 25 23  
## [176] 26 26 27 24 24 24 24 25 33 23 40 26 25 23 23 24 26 23 34 32 23 45 32 23 23  
## [201] 25 27 24 40 23 26 24 38 24 24 27 43 44 27 39 25 23 23 34 47 24 36 26 25 26  
## [226] 25 26 31 29 37 29 24 29

### **Z-Score**

Le Z-Score mesure combien d’écarts-types une donnée est éloignée de la moyenne. En général, si le Z-Score est supérieur à 3 ou inférieur à -3, cela peut être considéré comme une valeur aberrante.

# Calcul du Z-Score pour welfare\_sen2018  
z\_score <- (welfare\_sen2018$hhsize - mean(welfare\_sen2018$hhsize)) / sd(welfare\_sen2018$hhsize)  
  
# Identifier les valeurs aberrantes (Z-Score > 3 ou < -3)  
outliers\_z\_score <- welfare\_sen2018$hhsize[abs(z\_score) > 3]  
print(outliers\_z\_score)

## [1] 38 32 38 30 29 28 34 44 32 33 30 29 37 35 35 33 48 28 28 32 28 32 32 31 29  
## [26] 50 34 36 44 36 29 28 32 42 29 28 34 29 29 28 28 28 28 28 30 29 30 31 28 33  
## [51] 29 28 33 35 48 44 30 32 56 28 32 31 32 28 34 45 34 28 34 30 47 29 31 35 43  
## [76] 29 44 42 37 40 42 33 44 33 40 34 32 45 32 40 38 43 44 39 34 47 36 31 29 37  
## [101] 29 29

max\_outlier <- max(outliers)  
print(paste("La valeur maximale des valeurs aberrantes est ", max\_outlier))

## [1] "La valeur maximale des valeurs aberrantes est 56"

Bien que la taille maximale d’un ménage soit de 56, ce qui pourrait initialement être perçu comme une valeur aberrante, une vérification a révélé qu’il est en réalité possible qu’un ménage compte jusqu’à 56 personnes au Sénégal.

# **Section 4 : Traitement des variables**

## **Comparaison et harmonisation des variables entre les bases de 2018 et 2021**

# Créer un tableau récapitulatif des variables présentes dans vars\_2018\_base et vars\_2021\_base.  
  
recap <- tibble(# Créons une mini-base contenant la liste des variables.  
 Variable = unique(c(vars\_2018\_base, vars\_2021\_base)), # Garde les variables de façon unique, présentes dans au moins l'une des bases.  
 `Present en 2018` = ifelse(Variable %in% vars\_2018\_base, "✅", ""), # Cochez en vert si la variable est présente dans la base 2018.  
 `Present en 2021` = ifelse(Variable %in% vars\_2021\_base, "✅", "") # Cochez en vert si la variable est présente dans la base 2021.  
)  
  
# Afficher le tableau avec gtsummary  
recap %>%  
 gt::gt() %>%  
 gt::tab\_header(  
 title = "Comparaison des variables entre les bases Welfare du Sénégal ent 2018 et en 2021",  
 subtitle = "Variables communes et spécifiques à chaque année (2018 et 2021)"  
 )

Table 1: Comparaison des variables entre les bases Welfare du Sénégal ent 2018 et en 2021

Variables communes et spécifiques à chaque année (2018 et 2021)

| Variable | Present en 2018 | Present 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 comporte **35** variables sur **7156** individus, tandis que celle de 2021 a **47** variables sur **7120** individus. On constate qu’en tout, les deux bases possèdent **48** variables distinctes. Parmi celles-ci, **34** sont présentes dans les deux bases, **1** apparaissent uniquement en 2018 et **13** sont spécifiques à 2021.

**NB : La variable halfab de 2018 et la variable halfa de 2021 fournissent la même information, à savoir “Savoir lire et écrire du CM”. Cependant, en raison d’une différence d’orthographe, elles sont traitées comme distinctes dans le tableau.**

Pour remédier à cela, nous renommons celle de 2018 en halfa.

# Renommons halfab en halfa  
  
colnames(welfare\_sen2018)[colnames(welfare\_sen2018)=="halfab"] <- "halfa"  
  
# Ajoutons cette variables à la liste des variables communes.  
  
common\_vars\_2018\_2021 <- append(common\_vars\_2018\_2021,"halfa")  
  
common\_vars\_2018\_2021

## [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" "halfa"

## **Vérification de la labélisation**

Nous vérifions si les variables communes entre les bases welfare\_sen2018 et welfare\_sen2021 ont les mêmes labels. Pour ce faire, nous parcourons la liste des variables communes et récupérons les labels de chaque variable pour 2018 et 2021. Ensuite, nous comparons les labels afin d’identifier toute incohérence éventuelle.

variable\_label\_diff <- c() # créer une liste vide   
  
for (variable in common\_vars\_2018\_2021) { # parcourons les variables communes dans les deux bases welfare\_sen2018 et welfare\_sen2021.  
   
 if(labelled::is.labelled(welfare\_sen2018[[variable]])){ # Vérifier si la variable de la base 2018 est étiquetée.  
   
 value\_label\_base18 <- labelled::val\_labels(welfare\_sen2018[[variable]]) # Récupérer les labels de la variable dans la base 2018.  
  
 }else{  
 value\_label\_base18 <- NULL # Mettre vide si la variable en 2018 n'est pas étiquetée.  
 }  
   
 if(labelled::is.labelled(welfare\_sen2021[[variable]])){ # vérifier si la variable de la base 2021 est étiquetée.  
   
 value\_label\_base21 <- labelled::val\_labels(welfare\_sen2021[[variable]]) # recupérer les labels de la variable dans la base 2021.  
   
 }else{  
   
 value\_label\_base21 <- NULL # Mettre vide si la variable en 2021 n'est pas étiquetée.  
 }  
   
 if(!identical(value\_label\_base18, value\_label\_base21)){ # 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 créée.  
 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.

La taille de la liste contenant les variables dont les labels diffèrent d’une année à l’autre est égale à 6. Ainsi, nous avons**6 variable** qui ne sont pas étiquetées de la même façon. Traitons ces cas un par un afin d’assurer une labélisation identique.

### **1. Variable “zae”**

Elle représente la zone agro-écologique.

**Modalités de la variable “zae” en 2018**

labelled::val\_labels(welfare\_sen2018$zae)

## NULL

**Modalités de la variable “zae” en 2021**

labelled::val\_labels(welfare\_sen2021$zae)

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

La variable “zae” n’est pas étiquetée en 2018, mais l’est en 2021. Afin d’uniformiser, nous allons étiqueter la variable en 2018 en utilisant les mêmes labels qu’en 2021.

plot(welfare\_sen2021$zae)

**Recodons la variable “zae” en 2021**

Par ce processus, nous aurons les mêmes codifications en 2021 et en 2018, puis nous ajouterons les labels sur la variable de 2018. Nous constatons que chaque *zone agro-écologique* est liée à la région de résidence. En nous basant sur cette variable, nous pouvons recoder de manière conforme à celle de la *zone agro-écologique* de la base 2018.

**Distribution avant modification**

welfare\_sen2021 %>%  
 to\_factor() %>%  
 select(zae) %>%  
 tbl\_summary() %>%  
 modify\_header(label = "\*\*Répartition en pourcentage des zones agroécologiques par région au Sénégal en 2021\*\*")

| **Répartition en pourcentage des zones agroécologiques par région au Sénégal en 2021** | **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%) |
| *1*n (%) | |

**Recodage des modalités de la variable “zae” en 2021**

welfare\_sen2021$zae <- case\_when(  
 welfare\_sen2021$zae == 11 ~ 1, # 1 devient 11  
 welfare\_sen2021$zae == 9 ~ 5, # 5 devient 9  
 welfare\_sen2021$zae == 5 ~ 3, # 3 devient 5  
 welfare\_sen2021$zae == 3 ~ 2, # 2 devient 3  
 welfare\_sen2021$zae == 7 ~ 4, # 4 devient 7  
 welfare\_sen2021$zae == 1 ~ 6, # 6 devient 1  
 TRUE ~ welfare\_sen2021$zae # Les autres valeurs restent inchangées  
 )

**Labélisation des modalités de la variable “zae” en 2021**

Le recodage n’a pas modifié la labélisation. Nous l’adoptons à présent.

labelled::val\_labels(welfare\_sen2021$zae) <- c("Kédougou" = 6, "Saint-Louis-Matam"= 2, "Thies-Diourbel-Louga" = 3, "Kaolack-Fatick-Kaffrine" = 4, "Ziguinchor-Tamba-Kolda-Sédhiou" = 5,"Dakar" = 1)

**Vérification de la distribution après modification**

welfare\_sen2021 %>%  
 to\_factor() %>%  
 select(zae) %>%  
 tbl\_summary() %>%  
 modify\_header(label = "\*\*Répartition en pourcentage des zones agroécologiques par région au Sénégal en 2021\*\*")

| **Répartition en pourcentage des zones agroécologiques par région au Sénégal en 2021** | **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%) |
| *1*n (%) | |

Nous voyons que les valeurs sont restées inchangées. Donc, la labélisation a été correcte.

**Ajout des labels à la base welfare\_sen2018**

Maintenant que la codification est identique, nous ajoutons à présent les labels nécessaires à la base welfare\_sen2018.

labelled::val\_labels(welfare\_sen2018$zae) <- labelled::val\_labels(welfare\_sen2021$zae) # Affectons simplement les labels de la variable "zae" de la base welfare\_sen2021 à la base de welfare\_sen2018.  
print("Les labels de la variable `zae` de la base welfare\_sen2018 :")

## [1] "Les labels de la variable `zae` de la base welfare\_sen2018 :"

labelled::val\_labels(welfare\_sen2018$zae)

## Kédougou Saint-Louis-Matam   
## 6 2   
## Thies-Diourbel-Louga Kaolack-Fatick-Kaffrine   
## 3 4   
## Ziguinchor-Tamba-Kolda-Sédhiou Dakar   
## 5 1

print("Les labels de la variable `zae` de la base welfare\_sen2021 :")

## [1] "Les labels de la variable `zae` de la base welfare\_sen2021 :"

labelled::val\_labels(welfare\_sen2021$zae)

## Kédougou Saint-Louis-Matam   
## 6 2   
## Thies-Diourbel-Louga Kaolack-Fatick-Kaffrine   
## 3 4   
## Ziguinchor-Tamba-Kolda-Sédhiou Dakar   
## 5 1

### **2. Variable “hnation”**

**Modalités de la variable en 2018**

labelled::val\_labels(welfare\_sen2018$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 “hnation”en 2021**

labelled::val\_labels(welfare\_sen2021$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 constate que la présence de certains pays en 2021 qui ne figurent pas en 2018. Il s’agit du Liberia, Sierra Leone, Cap-Vert, Gambie, Ghana et Guinée, que nous regrouperons dans la variable Autre CEDEAO. De plus, nous réaménageons les modalités afin que cela soit cohérent dans les deux bases.

**Visualisation de la distribution avant la modification**

library(ggplot2)  
  
ggplot(data = welfare\_sen2021, aes(x = hnation, fill = is.na(hnation))) +   
 geom\_bar()

**Modification (recodage)**

library(dplyr)  
  
welfare\_sen2021 <- welfare\_sen2021 %>%  
 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  
 ))  
labelled::val\_labels(welfare\_sen2021$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

Maintenant que les codes sont uniformisés, nous allons ajouter les labels.

labelled::val\_labels(welfare\_sen2018$hnation) <- labelled::val\_labels(welfare\_sen2021$hnation) # Labélisons la variable hnation de la base 2018 en utilisant les labels de la base 2021.

### **3. Variable “hdiploma”**

**Modalités de la variable “hdiploma” de la base welfare\_sen2018**

labelled::val\_labels(welfare\_sen2018$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 “hdiploma” de la base welfare\_sen2021**

labelled::val\_labels(welfare\_sen2021$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 à noter concerne uniquement certains labels. Par exemple, **“cepe” en 2021 et “CEP/CFEE” en 2018**. Sinon, les modalités (codes) sont identiques. Pour résoudre cela, nous conserverons les labélisations de 2018 et mettrons à jour celles de 2021 en conséquence.

labelled::val\_labels(welfare\_sen2021$hdiploma) <- labelled::val\_labels(welfare\_sen2018$hdiploma)  
labelled::val\_labels(welfare\_sen2021$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

### **4. Variable “hactiv7j”**

**Modalités de la variable “hactiv7j” de la base welfare\_sen2018**

labelled::val\_labels(welfare\_sen2018$hactiv7j)

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

**Modalités de la variable “hactiv7j” de la base welfare\_sen2021**

labelled::val\_labels(welfare\_sen2021$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.

Ici, l’incohérence à noter réside dans le fait que les modalités et labels diffèrent légèrement. Par exemple, en 2018, **2 représente chômeur**, tandis qu’en 2021, **4** est utilisé pour **chômeur**. Pour résoudre cela, nous conserverons les labélisations de 2018 et mettrons à jour celles de 2021 en conséquence.

labelled::val\_labels(welfare\_sen2021$hactiv7j) <- labelled::val\_labels(welfare\_sen2018$hactiv7j)  
labelled::val\_labels(welfare\_sen2021$hactiv7j)

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

### **5. Variable “hbranch”**

**Modalités de la variable “hbranch” de la base welfare\_sen2018**

labelled::val\_labels(welfare\_sen2018$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 “hbranch” de la base welfare\_sen2021**

labelled::val\_labels(welfare\_sen2021$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.

Ici, l’incohérence à noter concerne uniquement certains labels. Par exemple, en 2021, **Elevage/syl./peche en 2021 et Elevage/peche en 2018**. Sinon, les modalités (codes) sont identiques. Pour résoudre cela, nous conserverons les labélisations de 2021 et mettrons à jour celles de 2018 en conséquence.

labelled::val\_labels(welfare\_sen2018$hbranch) <- labelled::val\_labels(welfare\_sen2021$hbranch)  
labelled::val\_labels(welfare\_sen2018$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

### **6. Variable “hcsp”**

**Modalités de la variable “hcsp” de la base welfare\_sen2018**

labelled::val\_labels(welfare\_sen2018$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 “hcsp” de la base welfare\_sen2021**

labelled::val\_labels(welfare\_sen2021$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 à noter concerne uniquement certains labels. Par exemple, en 2021, **Travailleur familial contribuant pour une entreprise familiale** et en 2018, **Travailleur familial contribuant à une entreprise familiale**. Sinon, les modalités (codes) sont identiques. Pour résoudre cela, nous conserverons les labélisations de 2021 et mettrons à jour celles de 2018 en conséquence.

labelled::val\_labels(welfare\_sen2018$hcsp) <- labelled::val\_labels(welfare\_sen2021$hcsp)  
labelled::val\_labels(welfare\_sen2018$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

### **7. Revérification de la labélisation**

variable\_label\_diff <- c() # créer une liste vide   
  
for (variable in common\_vars\_2018\_2021) { # parcourons les variables communes dans les deux bases welfare\_sen2018 et welfare\_sen2021.  
   
 if(labelled::is.labelled(welfare\_sen2018[[variable]])){ # Vérifier si la variable de la base 2018 est étiquetée.  
   
 value\_label\_base18 <- labelled::val\_labels(welfare\_sen2018[[variable]]) # Récupérer les labels de la variable dans la base 2018.  
  
 }else{  
 value\_label\_base18 <- NULL # Mettre vide si la variable en 2018 n'est pas étiquetée.  
 }  
   
 if(labelled::is.labelled(welfare\_sen2021[[variable]])){ # vérifier si la variable de la base 2021 est étiquetée.  
   
 value\_label\_base21 <- labelled::val\_labels(welfare\_sen2021[[variable]]) # recupérer les labels de la variable dans la base 2021.  
   
 }else{  
   
 value\_label\_base21 <- NULL # Mettre vide si la variable en 2021 n'est pas étiquetée.  
 }  
   
 if(!identical(value\_label\_base18, value\_label\_base21)){ # 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 créée.  
 print(variable)  
   
 }  
}

Maintenant, la taille de la liste contenant les variables dont les labels diffèrent d’une année à l’autre est égale à 0. Nous avons donc résolu ce problème.

## **Vérification des types de variables**

Toujours en considérant les variables communes, il est important de s’assurer que, d’une année à l’autre, chaque variable conserve le même type. Pour ce faire, nous utiliserons la fonction *class()*.

variable\_type\_diff <- c() # Création d'une liste vide pour stocker les variables avec des types différents.  
  
for (variable in common\_vars\_2018\_2021) { # Parcourir les variables communes dans les deux bases  
  
 type\_var\_2018 <- class(welfare\_sen2018[[variable]]) # Récupérer le type de la variable en 2018.  
 type\_var\_2021 <- class(welfare\_sen2021[[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 avons une seule variable (**zae**) dont le type varie de 2018 à 2021. La variable est de type **haven\_labelled, vctrs\_vctr, double** en 2018 et de type **haven\_labelled, vctrs\_vctr, double** en 2021. Nous allons maintenant uniformiser son type en numeric, afin d’éviter toute incohérence lors des opérations de fusion et de garantir l’intégrité des données.

welfare\_sen2021$zae <- as.numeric(welfare\_sen2021$zae) # Convertir la variable "zae" de 2021 en numerique.

## **Détection des ménages panels**

À présent, nous allons identifier les ménages présents en 2021 et vérifier lesquels étaient déjà présents en 2018, ainsi que ceux de 2018 qui ne figurent plus en 2021.

Pour ce faire, un extrait du document explicatif de l’EHCVM 2021 nous fournit des indications importantes : *<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.>*

Ainsi, l’utilisation des variables « grappe », « ménage », et « PanelHH » nous permettra d’effectuer le suivi des ménages en panel sur les deux vagues.

Ce qu’il faut retenir, c’est que pour chaque ménage de 2021, une variable permet de savoir si ce ménage était présent en 2018 ou non. Cette variable, **PanelHH**, est présente dans la base d’identification du ménage s00\_me\_sen\_2021. Nous allons donc extraire cette variable et l’ajouter à notre base welfare afin de pouvoir suivre l’appartenance des ménages au panel entre 2018 et 2021.

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

panel\_var <- haven::read\_dta("../donnees/s00\_me\_sen2021.dta") %>% # Importation de la base "s00\_me\_sen2021".  
 select(grappe,menage,PanelHH) # Sélection de la variable PanelHH et des identifiants du ménage (grappe, id).  
  
## Vérifions la taille de cette mini base  
  
length(panel\_var$menage)

## [1] 7120

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

Pour cela, nous effectuerons une jointure en utilisant l’id de chaque ménage comme identifiant principal. Cette jointure permettra d’ajouter la variable PanelHH à la base welfare, afin de déterminer, pour chaque ménage, s’il fait partie du panel en 2021, tout en reliant les informations aux données de 2018.

welfare\_2021\_ajuste <- dplyr::left\_join(welfare\_sen2021,panel\_var,by=c("grappe","menage"))

# **Section 5 : Fusion des deux bases**

Dans cette section, nous procéderons à une fusion optimale des deux bases de données, en privilégiant un empilement des données de 2018 et de 2021. Pour ce faire, au-delà des traitements déjà effectués, nous devrons nous assurer de certains aspects importants.

## **Ordre des variables**

Dans cette partie,comme toutes les variables de 2018 y sont en 2021, nous essaierons d’avoir dans chaque base, les variables communes au début et ordonnée dans le même sens. Ainsi, nous gardons Quant aux variables présentent seulement en 2021, nous les mettons en fin.

### **1. Extraction de l’ordre des colonnes communes**

# Extraire l’ordre des colonnes communes.  
common\_vars\_2018\_2021 <- names(welfare\_2021\_ajuste)[names(welfare\_2021\_ajuste) %in% names(welfare\_sen2018)]

### **2. Réorganisation des variables de la base 2018 selon l’ordre des variables de la base 2021**

# Réordonnons les variables dans la base 2018.  
welfare\_sen2018 <- welfare\_sen2018[, common\_vars\_2018\_2021]

### **3. Vérification de l’ordre des variables**

names(welfare\_sen2018) # Variables communes dans la base 2018.

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

names(welfare\_2021\_ajuste)[names(welfare\_2021\_ajuste) %in% names(welfare\_sen2018)] # Variables communes dans la base 2021.

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

Étant donné que l’ordre des variables est bien respecté, nous pouvons désormais procéder à l’empilement des bases.

### **4. Empilement**

Après avoir vérifié les différentes incohérences possibles, nous pouvons empiler les données en utilisant la fonction *bind\_rows*.

welfare\_combined <- bind\_rows(welfare\_sen2018,welfare\_2021\_ajuste) %>% # empiler les deux bases.  
 arrange(grappe,menage) # Trier les données en fonction de la grappe et du numéro du ménage."

### **5. Labélisation de la nouvelle base de données**

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

### **6. Extrait de la base de données fusionnée**

welfare\_combined[1:10,1:10] # Affichons les 10 premières lignes et les 10 premières variables de la base fusionnée.

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

# **Section 6 : Quelques analyses statistiques**

## **Analyse univarée**

Table1 <- welfare\_combined %>%  
 select(year, hgender, hage, hhsize, heduc, milieu, pcexp) %>% # Sélectionner les variables à afficher.  
 tbl\_summary(by = year, # Grouper les variables par année.  
   
 label = list( # Labéliser 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 capita (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 # Arrondir les nombres continus à 1 décimale.  
 ) %>%  
 modify\_header(label = "\*\*Profil sociodémographique et économique des ménages au Sénégal en 2018 et 2021\*\*") %>%  
 bold\_labels() # Mettre les labels en gras.  
  
Table1 # Afficher la table

| **Profil sociodémographique et économique des ménages au Sénégal en 2018 et 2021** | **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 capita (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 enquêtés, la proportion de **femmes chefs de ménage** a légèrement augmenté, passant de **26% à 28%**, tandis que celle des **hommes** a diminué, passant de **74% à 72%**. **L’âge moyen des chefs de ménage** est passé de **51,5 à 54,1 ans**, ce qui reflète 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 petits.

Concernant **l’éducation**, la proportion des chefs de ménage n’ayant **jamais été scolarisés a légèrement augmenté, passant de 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 milieu rural (45%)** est restée inchangée. Enfin, **la dépense annuelle moyenne par habitant a légèrement augmenté, passant de 615 630 FCFA en 2018 à 621 198 FCFA en 2021**, avec cependant une grande variabilité entre les ménages.

## **Analyse bivariée**

L’analyse bivariée suivante examine la relation entre le sexe du chef de ménage et le milieu de résidence (urbain ou rural) dans les données de 2021.

### **1. Tableau croisé de fréquences et test du Chi-carré**

# Charger les bibliothèques nécessaires  
library(dplyr)  
library(knitr) # Pour afficher un tableau avec un titre  
library(gt) # Optionnel pour un affichage amélioré  
  
# 1. Tableau croisé de fréquences  
tableau\_croise <- table(welfare\_combined$hgender, welfare\_combined$milieu)  
  
# Afficher le tableau croisé avec un titre  
cat("### Tableau croisé : Sexe du chef de ménage et Milieu de résidence\n")

## ### Tableau croisé : Sexe du chef de ménage et Milieu de résidence

kable(tableau\_croise, caption = "Répartition du sexe du chef de ménage selon le milieu de résidence")

Répartition du sexe du chef de ménage selon le milieu de résidence

|  | Urbain | Rural |
| --- | --- | --- |
| Masculin | 5171 | 5204 |
| Féminin | 2692 | 1209 |

# 2. Tableau croisé avec pourcentages  
tableau\_croise\_percentages <- prop.table(tableau\_croise) \* 100  
  
cat("\n### Tableau croisé en pourcentage : Sexe du chef de ménage et Milieu de résidence\n")

##   
## ### Tableau croisé en pourcentage : Sexe du chef de ménage et Milieu de résidence

kable(tableau\_croise\_percentages, caption = "Répartition en pourcentage du sexe du chef de ménage selon le milieu de résidence")

Répartition en pourcentage du sexe du chef de ménage selon le milieu de résidence

|  | Urbain | Rural |
| --- | --- | --- |
| Masculin | 36.22163 | 36.452788 |
| Féminin | 18.85682 | 8.468759 |

# 3. Test du Chi-carré  
chi\_test <- chisq.test(tableau\_croise)  
cat("\n### Résultat du test du Chi-carré\n")

##   
## ### Résultat du test du Chi-carré

print(chi\_test)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: tableau\_croise  
## X-squared = 420.17, df = 1, p-value < 2.2e-16

Le tableau croisé des fréquences montre la répartition des chefs de ménage selon leur sexe et leur milieu de résidence. En 2021, la majorité des **chefs de ménage masculins** vivent en milieu **rural (36,45%)** et **urbain (36,22%)**. En revanche, **les chefs de ménage féminins** sont proportionnellement plus nombreux en milieu **urbain (18,86%)** qu’en milieu **rural (8,47%)**. Le test du Chi-carré révèle une association statistiquement significative entre le sexe du chef de ménage et le milieu de résidence, avec **une p-value inférieure à 2.2e-16**. Cela indique qu’il existe **une relation forte et significative entre ces deux variables**.

Bref, **le sexe du chef de ménage semble influencer son lieu de résidence, les femmes étant plus présentes dans les zones urbaines, tandis que les hommes sont plus répartis entre milieu urbain et rural**.

### **2. Nuage de points**

# Charger les bibliothèques nécessaires  
library(ggplot2)  
  
# 3. Nuage de points (pour deux variables continues)  
# Supposons que les variables continues soient 'hage' et 'pcexp'  
# Remplacez 'hage' et 'pcexp' par les noms réels de vos variables continues  
  
cat("\nNuage de points entre hage et pcexp:\n")

##   
## Nuage de points entre hage et pcexp:

ggplot(welfare\_combined, aes(x = hage, y = pcexp)) +  
 geom\_point(color = "blue") +  
 labs(x = "Âge du chef de ménage", y = "Dépense par habitant (FCFA)",   
 title = "Nuage de points entre Âge du chef de ménage et Dépense par habitant") +  
 theme\_minimal()

# **Section 9 : Vérification de la base finale**

## **Vérification de la structure de la base**

cat("Dimensions de la base fusionnée :", dim(welfare\_combined), "\n")

## Dimensions de la base fusionnée : 14276 48

cat("Aperçu des premières lignes :\n")

## Aperçu des premières lignes :

print(head(welfare\_combined))

## # A tibble: 6 × 48  
## grappe menage country year hhid vague zae region milieu hhweight hhsize  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <fct> <fct> <fct> <dbl> <dbl>  
## 1 1 1 SEN 2018 1001 1 Dakar dakar Urbain 1750. 2  
## 2 1 2 SEN 2018 1002 1 Dakar dakar Urbain 1750. 2  
## 3 1 3 SEN 2018 1003 1 Dakar dakar Urbain 1750. 1  
## 4 2 1 SEN 2018 2001 2 Dakar dakar Urbain 266. 10  
## 5 2 1 SEN 2021 201 2 Dakar dakar Urbain 290. 9  
## 6 2 2 SEN 2018 2002 2 Dakar dakar Urbain 266. 6  
## # ℹ 37 more variables: eqadu1 <dbl>, eqadu2 <dbl>, hgender <fct>, hage <dbl>,  
## # hmstat <fct>, hreligion <fct>, hnation <fct>, halfa <fct>, heduc <fct>,  
## # hdiploma <fct>, hhandig <fct>, hactiv7j <fct>, hactiv12m <fct>,  
## # hbranch <fct>, hsectins <fct>, hcsp <fct>, dali <dbl>, dnal <dbl>,  
## # dtot <dbl>, pcexp <dbl>, zzae <dbl>, zref <dbl>, def\_spa <dbl>,  
## # def\_temp <dbl>, month <date>, hethnie <fct>, halfa2 <fct>,  
## # def\_temp\_prix2021m11 <dbl>, def\_temp\_cpi <dbl>, def\_temp\_adj <dbl>, …

cat("Types de variables :\n")

## Types de variables :

print(str(welfare\_combined))

## tibble [14,276 × 48] (S3: tbl\_df/tbl/data.frame)  
## $ grappe : num [1:14276] 1 1 1 2 2 2 2 2 2 2 ...  
## $ menage : num [1:14276] 1 2 3 1 1 2 3 3 4 4 ...  
## $ country : chr [1:14276] "SEN" "SEN" "SEN" "SEN" ...  
## $ year : num [1:14276] 2018 2018 2018 2018 2021 ...  
## $ hhid : num [1:14276] 1001 1002 1003 2001 201 ...  
## $ vague : num [1:14276] 1 1 1 2 2 2 2 2 2 2 ...  
## $ zae : Factor w/ 6 levels "Kédougou","Saint-Louis-Matam",..: 6 6 6 6 6 6 6 6 6 6 ...  
## ..- attr(\*, "label")= chr "Zone agroecologique"  
## $ region : Factor w/ 14 levels "dakar","ziguinchor",..: 1 1 1 1 1 1 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "Region residence"  
## $ milieu : Factor w/ 2 levels "Urbain","Rural": 1 1 1 1 1 1 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "Milieu residence"  
## $ hhweight : num [1:14276] 1750 1750 1750 266 290 ...  
## $ hhsize : num [1:14276] 2 2 1 10 9 6 4 3 3 3 ...  
## $ eqadu1 : num [1:14276] 1.32 1.45 1 7.66 6.76 ...  
## $ eqadu2 : num [1:14276] 1.61 1.61 1 5.54 5.02 ...  
## $ hgender : Factor w/ 2 levels "Masculin","Féminin": 2 1 1 1 1 1 2 2 1 1 ...  
## ..- attr(\*, "label")= chr "Genre du CM"  
## $ hage : num [1:14276] 59 59 27 85 88 40 68 47 31 34 ...  
## $ hmstat : Factor w/ 7 levels "Célibataire",..: 1 2 1 2 2 2 5 3 1 1 ...  
## ..- attr(\*, "label")= chr "Situation famille du CM"  
## $ hreligion : Factor w/ 5 levels "Musulman","Chrétien",..: 1 2 2 1 1 1 1 1 2 2 ...  
## ..- attr(\*, "label")= chr "Religion du CM"  
## $ hnation : Factor w/ 17 levels "Bénin","Burkina Faso",..: 7 12 7 7 7 7 7 7 11 11 ...  
## ..- attr(\*, "label")= chr "Nationalite du CM"  
## $ halfa : Factor w/ 2 levels "Non","Oui": 2 2 2 2 2 1 2 2 2 2 ...  
## ..- attr(\*, "label")= chr "Alphabetisation du CM"  
## $ heduc : Factor w/ 9 levels "Aucun","Maternelle",..: 3 9 9 3 3 1 1 4 9 9 ...  
## ..- attr(\*, "label")= chr "Education du CM"  
## $ hdiploma : Factor w/ 11 levels "Aucun","CEP/CFEE",..: 2 7 11 1 1 1 1 2 11 11 ...  
## ..- attr(\*, "label")= chr "Diplome du CM"  
## $ hhandig : Factor w/ 2 levels "Non","Oui": 1 1 1 1 1 1 1 1 1 1 ...  
## ..- attr(\*, "label")= chr "Handicap majeur CM"  
## $ hactiv7j : Factor w/ 6 levels "Occupe","Chomeur",..: 1 1 1 5 1 1 5 1 1 5 ...  
## ..- attr(\*, "label")= chr "Activite 7 jours du CM"  
## $ hactiv12m : Factor w/ 4 levels "Occupe","Trav. fam.",..: 1 1 1 3 1 1 3 1 1 3 ...  
## ..- attr(\*, "label")= chr "Activite 12 mois du CM"  
## $ hbranch : Factor w/ 14 levels "Agriculture",..: 6 6 11 NA 2 6 NA 11 9 NA ...  
## ..- attr(\*, "label")= chr "Branche activite du CM"  
## $ hsectins : Factor w/ 6 levels "Etat/Collectivités locales",..: 3 3 1 NA 3 3 NA 5 1 NA ...  
## ..- attr(\*, "label")= chr "Secteur instit. du CM"  
## $ hcsp : Factor w/ 10 levels "Cadre supérieur",..: 4 1 7 NA 9 4 NA 5 6 NA ...  
## ..- attr(\*, "label")= chr "CSP du CM"  
## $ dali : num [1:14276] 1978488 1886313 387989 6568518 3150153 ...  
## $ dnal : num [1:14276] 13139233 8894985 5055663 4889539 3046992 ...  
## $ dtot : num [1:14276] 15117721 10781298 5443652 11458057 6197145 ...  
## $ pcexp : num [1:14276] 6440507 4593088 4638249 976281 600079 ...  
## $ zzae : num [1:14276] 391340 391340 391340 391340 424179 ...  
## $ zref : num [1:14276] 333441 333441 333441 333441 369666 ...  
## $ def\_spa : num [1:14276] 1.17 1.17 1.17 1.17 1.15 ...  
## $ def\_temp : num [1:14276] 1.015 1.012 1.015 0.992 0.991 ...  
## $ month : Date[1:14276], format: NA NA ...  
## $ hethnie : Factor w/ 13 levels "Wolof/Lébou",..: NA NA NA NA 8 NA NA 6 NA NA ...  
## ..- attr(\*, "label")= chr "Ethnie du CM"  
## $ halfa2 : Factor w/ 2 levels "Non","Oui": NA NA NA NA 2 NA NA 2 NA 2 ...  
## ..- attr(\*, "label")= chr "Alpha. lire/ecr./comp. CM"  
## $ def\_temp\_prix2021m11: num [1:14276] NA NA NA NA 1.03 ...  
## $ def\_temp\_cpi : num [1:14276] NA NA NA NA 1 ...  
## $ def\_temp\_adj : num [1:14276] NA NA NA NA 1.01 ...  
## $ zali0 : num [1:14276] NA NA NA NA 196233 ...  
## $ dtet : num [1:14276] NA NA NA NA 694665 ...  
## $ monthly\_cpi : num [1:14276] NA NA NA NA 120 ...  
## $ cpi2017 : num [1:14276] NA NA NA NA 1.1 ...  
## $ icp2017 : num [1:14276] NA NA NA NA 239 ...  
## $ dollars : num [1:14276] NA NA NA NA 6.95 ...  
## $ PanelHH : Factor w/ 2 levels "Nouveau Ménage",..: NA NA NA NA 2 NA NA 1 NA 2 ...  
## ..- attr(\*, "label")= chr "Type de ménage"  
## NULL

## **Vérification des valeurs manquantes**

cat("Valeurs manquantes par variable :\n")

## Valeurs manquantes par variable :

print(colSums(is.na(welfare\_combined)))

## grappe menage country   
## 0 0 0   
## year hhid vague   
## 0 0 0   
## zae region milieu   
## 0 0 0   
## hhweight hhsize eqadu1   
## 0 0 0   
## eqadu2 hgender hage   
## 0 0 0   
## hmstat hreligion hnation   
## 2 0 0   
## halfa heduc hdiploma   
## 0 0 0   
## hhandig hactiv7j hactiv12m   
## 0 0 0   
## hbranch hsectins hcsp   
## 3560 3081 3048   
## dali dnal dtot   
## 0 0 0   
## pcexp zzae zref   
## 0 0 0   
## def\_spa def\_temp month   
## 0 0 7156   
## hethnie halfa2 def\_temp\_prix2021m11   
## 7238 7156 7156   
## def\_temp\_cpi def\_temp\_adj zali0   
## 7156 7156 7156   
## dtet monthly\_cpi cpi2017   
## 7156 7156 7156   
## icp2017 dollars PanelHH   
## 7156 7156 7156

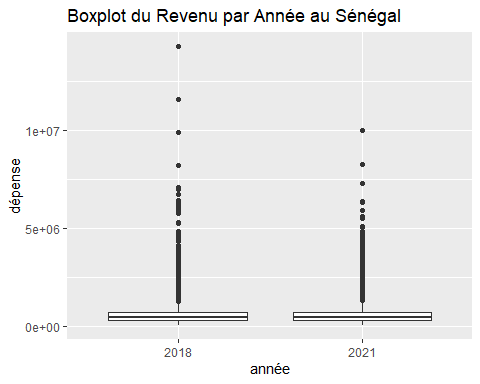
## **Vérification des doublons**

duplicate\_count <- sum(duplicated(welfare\_combined))  
cat("Nombre de doublons :", duplicate\_count, "\n")

## Nombre de doublons : 0

## **Vérification des valeurs aberrantes (Ex: pcexp)**

ggplot(welfare\_combined, aes(x = as.factor(year), y = pcexp)) +  
 geom\_boxplot() +   
 labs(title = "Boxplot du Revenu par Année au Sénégal", x = "année", y = "dépense")



## **Vérification de la répartition des données par année**

cat("Effectif des observations par année :\n")

## Effectif des observations par année :

print(table(welfare\_combined$year))

##   
## 2018 2021   
## 7156 7120

## **Résumé statistique de la base**

cat("Résumé statistique :\n")

## Résumé statistique :

print(summary(welfare\_combined))

## grappe menage country year   
## Min. : 1.0 Min. : 1.000 Length:14276 Min. :2018   
## 1st Qu.:151.0 1st Qu.: 4.000 Class :character 1st Qu.:2018   
## Median :300.0 Median : 7.000 Mode :character Median :2018   
## Mean :300.3 Mean : 6.838 Mean :2019   
## 3rd Qu.:450.0 3rd Qu.:10.000 3rd Qu.:2021   
## Max. :598.0 Max. :19.000 Max. :2021   
##   
## hhid vague zae   
## Min. : 201 Min. :1.000 Kédougou : 908   
## 1st Qu.: 27511 1st Qu.:1.000 Saint-Louis-Matam :1823   
## Median : 54707 Median :2.000 Thies-Diourbel-Louga :3201   
## Mean :165348 Mean :1.502 Kaolack-Fatick-Kaffrine :2827   
## 3rd Qu.:300011 3rd Qu.:2.000 Ziguinchor-Tamba-Kolda-Sédhiou:3492   
## Max. :598012 Max. :2.000 Dakar :2025   
##   
## region milieu hhweight hhsize   
## dakar :2025 Urbain:7863 Min. : 15.29 Min. : 1.000   
## thies :1137 Rural :6413 1st Qu.: 123.53 1st Qu.: 5.000   
## diourbel :1104 Median : 212.14 Median : 8.000   
## kaolack :1054 Mean : 273.85 Mean : 8.994   
## SAINT-LOUIS:1007 3rd Qu.: 346.76 3rd Qu.:11.000   
## louga : 960 Max. :3081.51 Max. :56.000   
## (Other) :6989   
## eqadu1 eqadu2 hgender hage   
## Min. : 0.660 Min. : 1.000 Masculin:10375 Min. : 16.00   
## 1st Qu.: 3.959 1st Qu.: 3.088 Féminin : 3901 1st Qu.: 42.00   
## Median : 5.885 Median : 4.257 Median : 52.00   
## Mean : 6.728 Mean : 4.718 Mean : 52.78   
## 3rd Qu.: 8.531 3rd Qu.: 5.836 3rd Qu.: 63.00   
## Max. :41.240 Max. :23.706 Max. :101.00   
##   
## hmstat hreligion hnation   
## Marié(e) monogame:7786 Musulman :13467 Guinee :14061   
## Marié(e) polygame:3780 Chrétien : 777 Mali : 81   
## Veuf(ve) :1933 Animiste : 21 Gambie : 45   
## Célibataire : 389 Autre Réligion: 7 Niger : 44   
## Divorcé(e) : 337 Sans Réligion : 4 Cote d'ivoire: 26   
## (Other) : 49 Nigeria : 6   
## NA's : 2 (Other) : 13   
## halfa heduc hdiploma hhandig   
## Non:7265 Aucun :9727 Aucun :11469 Non:12991   
## Oui:7011 Primaire :2110 CEP/CFEE : 1170 Oui: 1285   
## Second. gl 1 :1046 BEPC/BFEM: 676   
## Superieur : 629 bac : 304   
## Second. gl 2 : 545 Licence : 184   
## Postsecondaire: 114 Maitrise : 118   
## (Other) : 105 (Other) : 355   
## hactiv7j hactiv12m hbranch   
## Occupe :10540 Occupe :11098 Agriculture :2729   
## Chomeur : 49 Trav. fam. : 130 Commerce :1998   
## TF cherchant emploi: 65 Non occupe : 3048 Autr. indust. :1140   
## TF cherchant pas : 94 Moins de 5 ans: 0 Services perso. :1023   
## Inactif : 3528 Elevage/syl./peche: 897   
## Moins de 5 ans : 0 (Other) :2929   
## NA's :3560   
## hsectins   
## Etat/Collectivités locales : 496   
## Entreprise publique/ parapublique : 331   
## Entreprise Privée :10123   
## Entreprise associative : 121   
## Ménage comme employeur de personnel domestique: 112   
## Organisme international /Ambassade : 12   
## NA's : 3081   
## hcsp dali dnal   
## Travailleur pour compte propre :7993 Min. : 64205 Min. : 27585   
## Ouvrier ou employé qualifié : 839 1st Qu.: 1227822 1st Qu.: 983135   
## Ouvrier ou employé non qualifié: 822 Median : 1847097 Median : 1609784   
## Cadre moyen/agent de maîtrise : 595 Mean : 2169794 Mean : 2148875   
## Manœuvre, aide ménagère : 304 3rd Qu.: 2725284 3rd Qu.: 2652194   
## (Other) : 675 Max. :31295272 Max. :221582921   
## NA's :3048   
## dtot pcexp zzae zref   
## Min. : 235210 Min. : 57610 Min. :296311 Min. :333441   
## 1st Qu.: 2350300 1st Qu.: 313625 1st Qu.:326012 1st Qu.:333441   
## Median : 3546130 Median : 456803 Median :335734 Median :333441   
## Mean : 4318669 Mean : 618407 Mean :351816 Mean :351507   
## 3rd Qu.: 5378076 3rd Qu.: 712685 3rd Qu.:375492 3rd Qu.:369666   
## Max. :227607152 Max. :14286279 Max. :424179 Max. :369666   
##   
## def\_spa def\_temp month hethnie   
## Min. :0.8886 Min. :0.9455 Min. :2021-11-01 Poular :2383   
## 1st Qu.:0.9169 1st Qu.:0.9916 1st Qu.:2021-12-01 Wolof/Lébou :2269   
## Median :0.9957 Median :0.9955 Median :2022-04-01 Sérère : 872   
## Mean :1.0007 Mean :0.9998 Mean :2022-02-16 Mandingue/Socé: 445   
## 3rd Qu.:1.0430 3rd Qu.:1.0118 3rd Qu.:2022-05-01 Diola : 409   
## Max. :1.1736 Max. :1.0590 Max. :2022-07-01 (Other) : 660   
## NA's :7156 NA's :7238   
## halfa2 def\_temp\_prix2021m11 def\_temp\_cpi def\_temp\_adj   
## Non :3598 Min. :1.000 Min. :0.974 Min. :0.961   
## Oui :3522 1st Qu.:1.001 1st Qu.:0.975 1st Qu.:0.999   
## NA's:7156 Median :1.020 Median :0.993 Median :1.011   
## Mean :1.020 Mean :0.993 Mean :1.014   
## 3rd Qu.:1.027 3rd Qu.:1.001 3rd Qu.:1.036   
## Max. :1.087 Max. :1.059 Max. :1.076   
## NA's :7156 NA's :7156 NA's :7156   
## zali0 dtet monthly\_cpi cpi2017   
## Min. :196233 Min. : 50712 Min. :117.2 Min. :1.097   
## 1st Qu.:196233 1st Qu.: 317994 1st Qu.:117.4 1st Qu.:1.097   
## Median :196233 Median : 469028 Median :119.6 Median :1.097   
## Mean :196233 Mean : 641538 Mean :119.6 Mean :1.097   
## 3rd Qu.:196233 3rd Qu.: 746041 3rd Qu.:120.4 3rd Qu.:1.097   
## Max. :196233 Max. :10366096 Max. :127.4 Max. :1.097   
## NA's :7156 NA's :7156 NA's :7156 NA's :7156   
## icp2017 dollars PanelHH   
## Min. :238.6 Min. : 0.538 Nouveau Ménage: 993   
## 1st Qu.:238.6 1st Qu.: 3.247 Menage Panel :6127   
## Median :238.6 Median : 4.792 NA's :7156   
## Mean :238.6 Mean : 6.553   
## 3rd Qu.:238.6 3rd Qu.: 7.613   
## Max. :238.6 Max. :102.581   
## NA's :7156 NA's :7156

cat("✅ Vérification terminée !\n")

## ✅ Vérification terminée !

# **Section 10 : Modélisation**

Supposons que la variable dépendante soit pcexp (dépenses de consommation par tête), et que vous souhaitez analyser comment des variables comme hhsize (taille du ménage), hage (âge du chef de ménage), heduc (niveau d’éducation), hmstat (statut marital), etc. influencent cette variable.

## **Modélisation OLS - Régression Linéaire**

# Charger les bibliothèques nécessaires  
library(lmtest) # lmtest permet de réaliser des tests statistiques sur des modèles de régression, tels que les tests de significativité des coefficients et la détection de l'hétéroscédasticité ou de l'autocorrélation des résidus.  
  
library(sandwich) # sandwich fournit des estimateurs robustes des erreurs standards, permettant de corriger les problèmes d'hétéroscédasticité et d'autocorrélation dans les modèles de régression.  
  
# Définition du modèle OLS : pcexp expliqué par des variables socio-économiques  
model\_ols <- lm(pcexp ~ hhsize + hgender + hage + heduc + hmstat + region + year,   
 data = welfare\_combined)  
  
# Résumé des résultats  
summary(model\_ols)

##   
## Call:  
## lm(formula = pcexp ~ hhsize + hgender + hage + heduc + hmstat +   
## region + year, data = welfare\_combined)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1524793 -217650 -67828 115187 14142068   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6462922.5 5347782.3 1.209 0.22687   
## hhsize -32439.0 765.9 -42.353 < 2e-16 \*\*\*  
## hgenderFéminin 57643.6 11728.2 4.915 8.98e-07 \*\*\*  
## hage 2842.0 319.9 8.884 < 2e-16 \*\*\*  
## heducMaternelle 172707.3 272155.1 0.635 0.52571   
## heducPrimaire 95618.6 11676.1 8.189 2.85e-16 \*\*\*  
## heducSecond. gl 1 236819.4 15779.3 15.008 < 2e-16 \*\*\*  
## heducSecond. tech. 1 364645.3 62756.9 5.810 6.36e-09 \*\*\*  
## heducSecond. gl 2 361932.6 21116.8 17.140 < 2e-16 \*\*\*  
## heducSecond. tech. 2 393201.2 70536.6 5.574 2.53e-08 \*\*\*  
## heducPostsecondaire 531604.3 44505.1 11.945 < 2e-16 \*\*\*  
## heducSuperieur 856177.7 20078.4 42.642 < 2e-16 \*\*\*  
## hmstatMarié(e) monogame -415005.5 25051.9 -16.566 < 2e-16 \*\*\*  
## hmstatMarié(e) polygame -350263.3 26346.2 -13.295 < 2e-16 \*\*\*  
## hmstatUnion libre -463444.4 236971.3 -1.956 0.05052 .   
## hmstatVeuf(ve) -400404.5 29254.2 -13.687 < 2e-16 \*\*\*  
## hmstatDivorcé(e) -273286.9 35982.6 -7.595 3.27e-14 \*\*\*  
## hmstatSéparé(e) -219156.2 74509.8 -2.941 0.00327 \*\*   
## regionziguinchor -459227.8 18629.9 -24.650 < 2e-16 \*\*\*  
## regiondiourbel -287894.4 18175.3 -15.840 < 2e-16 \*\*\*  
## regionSAINT-LOUIS -256297.3 18491.1 -13.861 < 2e-16 \*\*\*  
## regiontambacounda -306122.1 19767.5 -15.486 < 2e-16 \*\*\*  
## regionkaolack -241569.2 18353.0 -13.162 < 2e-16 \*\*\*  
## regionthies -268763.7 17714.5 -15.172 < 2e-16 \*\*\*  
## regionlouga -253979.9 18862.2 -13.465 < 2e-16 \*\*\*  
## regionfatick -330999.4 18992.4 -17.428 < 2e-16 \*\*\*  
## regionkolda -382812.1 19589.6 -19.542 < 2e-16 \*\*\*  
## regionmatam -290415.6 20149.0 -14.413 < 2e-16 \*\*\*  
## regionkaffrine -281516.2 19811.3 -14.210 < 2e-16 \*\*\*  
## regionkedougou -372052.6 19181.5 -19.396 < 2e-16 \*\*\*  
## regionsedhiou -378200.2 19945.0 -18.962 < 2e-16 \*\*\*  
## year -2554.7 2648.7 -0.964 0.33482   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 470800 on 14242 degrees of freedom  
## (2 observations deleted due to missingness)  
## Multiple R-squared: 0.3482, Adjusted R-squared: 0.3468   
## F-statistic: 245.5 on 31 and 14242 DF, p-value: < 2.2e-16

## **Vérification des Hypothèses**

### **1. Signification des Coefficients**

Regardons la p-value des variables explicatives pour déterminer si elles sont significativement différentes de zéro (si p-value < 0.05, elles sont significatives).

### **2. Test de l’Hétéroscédasticité**

Vérifions si l’hypothèse d’homoscédasticité (variances constantes) est respectée avec le test de Breusch-Pagan. Si l’hétéroscédasticité est présente, vous devrez peut-être ajuster les erreurs standards.

# Test de Breusch-Pagan pour vérifier l'hétéroscédasticité  
bptest(model\_ols)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_ols  
## BP = 292.15, df = 31, p-value < 2.2e-16

Si le test de Breusch-Pagan retourne une p-value faible (p < 0.05), cela suggère que l’hétéroscédasticité est présente, et il faudra peut-être utiliser des erreurs standards robustes.

### **3. Estimation avec Erreurs Standard Robustes**

Si l’hétéroscédasticité est présente, vous pouvez calculer les erreurs standard robustes.

# Calcul des erreurs standard robustes  
coeftest(model\_ols, vcov = vcovHC(model\_ols, type = "HC3"))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6462922.47 5296342.92 1.2203 0.222386   
## hhsize -32439.02 958.60 -33.8400 < 2.2e-16 \*\*\*  
## hgenderFéminin 57643.64 13164.27 4.3788 1.202e-05 \*\*\*  
## hage 2842.01 364.17 7.8042 6.404e-15 \*\*\*  
## heducMaternelle 172707.32 86165.58 2.0044 0.045050 \*   
## heducPrimaire 95618.61 9764.38 9.7926 < 2.2e-16 \*\*\*  
## heducSecond. gl 1 236819.43 19055.35 12.4280 < 2.2e-16 \*\*\*  
## heducSecond. tech. 1 364645.29 117022.64 3.1160 0.001837 \*\*   
## heducSecond. gl 2 361932.61 27012.93 13.3985 < 2.2e-16 \*\*\*  
## heducSecond. tech. 2 393201.17 91694.52 4.2882 1.813e-05 \*\*\*  
## heducPostsecondaire 531604.29 73158.94 7.2664 3.883e-13 \*\*\*  
## heducSuperieur 856177.74 48528.05 17.6429 < 2.2e-16 \*\*\*  
## hmstatMarié(e) monogame -415005.49 44632.05 -9.2984 < 2.2e-16 \*\*\*  
## hmstatMarié(e) polygame -350263.31 45652.31 -7.6724 1.797e-14 \*\*\*  
## hmstatUnion libre -463444.41 179923.49 -2.5758 0.010011 \*   
## hmstatVeuf(ve) -400404.53 47929.99 -8.3539 < 2.2e-16 \*\*\*  
## hmstatDivorcé(e) -273286.94 57234.36 -4.7749 1.816e-06 \*\*\*  
## hmstatSéparé(e) -219156.22 105855.84 -2.0703 0.038440 \*   
## regionziguinchor -459227.76 20371.79 -22.5423 < 2.2e-16 \*\*\*  
## regiondiourbel -287894.35 18737.10 -15.3649 < 2.2e-16 \*\*\*  
## regionSAINT-LOUIS -256297.29 20805.09 -12.3190 < 2.2e-16 \*\*\*  
## regiontambacounda -306122.07 24358.91 -12.5671 < 2.2e-16 \*\*\*  
## regionkaolack -241569.16 20336.90 -11.8784 < 2.2e-16 \*\*\*  
## regionthies -268763.66 20588.81 -13.0539 < 2.2e-16 \*\*\*  
## regionlouga -253979.92 19884.81 -12.7726 < 2.2e-16 \*\*\*  
## regionfatick -330999.37 20461.67 -16.1766 < 2.2e-16 \*\*\*  
## regionkolda -382812.10 19590.10 -19.5411 < 2.2e-16 \*\*\*  
## regionmatam -290415.62 19565.19 -14.8435 < 2.2e-16 \*\*\*  
## regionkaffrine -281516.16 23832.97 -11.8120 < 2.2e-16 \*\*\*  
## regionkedougou -372052.64 20626.26 -18.0378 < 2.2e-16 \*\*\*  
## regionsedhiou -378200.25 20634.46 -18.3286 < 2.2e-16 \*\*\*  
## year -2554.66 2622.12 -0.9743 0.329938   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##**Exportation de la nouvelle base combinée**

Après avoir combiné les deux bases, nous l’exporterons au format **.DTA** et la renommerons hcvm\_welfare\_sen2018\_2021 afin de pouvoir l’utiliser à tout moment.

# Exporter la base au format .dta  
write\_dta(welfare\_combined, "D:/TP9\_Paulinah\_RASAMOELINA/donnees/ehcvm\_welfare\_sen2018\_2021.dta")