TP9 : Merge des bases welfare 2018 et 2021

FALL Ndeye Ramatoulaye Ndoye

2025-03-27

#TP9: Merge des bases welfare des EHCVM 2018 et 2021 du Sénégal en une seule base

L’objectif de ce travail pratique (TP) est de fusionner les bases de données welfare des Enquêtes sur les Conditions de Vie des Ménages (EHCVM) 2018 et 2021 du Sénégal, afin de créer une base de données unifiée permettant des analyses comparatives. Nous allons explorer et préparer les données pour d’éventuelles analyses statistiques, en traitant des aspects comme les valeurs manquantes, les doublons, et les valeurs aberrantes. À la fin de ce TP, la base fusionnée sera nettoyée et prête pour des analyses plus poussées, comme des études sur les conditions de vie des ménages au Sénégal.

# 1. Installation et chargement des packages

Cette section assure la disponibilité des bibliothèques nécessaires.

Packages <- c("readr", "utils", "dplyr","labelled", "gtsummary","haven", "ggplot2", "tidyr", "plotly")  
  
for (x in Packages) {  
 if (!requireNamespace(x, quietly = TRUE)) {  
 install.packages(x)  
 }  
 library(x, character.only = TRUE)  
}

# 2. Importation des bases de données welfare

getwd()

## [1] "C:/Users/ANSD/Documents/TP9\_FALL\_Ndeye Ramatoulaye Ndoye/TP9\_way/Scripts"

welfare2018 <- haven::read\_dta("../Données/ehcvm\_welfare\_sen2018.dta") #importation base welfare 2018  
welfare2021 <- haven::read\_dta("../Données/ehcvm\_welfare\_sen2021.dta") #importation base welfare 2021

# 3. Visualisation des données :

Cette partie va nous permettre de voir la structure des données et d’avoir un aperçu sur les variables. Nous allons examiner les premières lignes de chaque fichier afin de nous assurer qu’ils ont été correctement lus.

# Aperçu des premières lignes pour chaque base  
  
# Cette étape nous permet de vérifier rapidement le format des données  
head(welfare2018)

## # 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>

head(welfare2021)

## # 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 des dimensions des deux bases  
# On vérifie combien de lignes et de colonnes chaque base contient  
dim(welfare2018)

## [1] 7156 35

dim(welfare2021)

## [1] 7120 47

# Informations générales sur les colonnes/variables  
# On utilise str() pour examiner les types de données et les noms de variables  
str(welfare2018)

## 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"

str(welfare2021)

## 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"

colnames(welfare2018)

## [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"

colnames(welfare2021)

## [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"

# 4. Vérification de la qualité des données :

Nous allons maintenant vérifier la qualité des données, où nous allons analyser la présence de valeurs manquantes et de doublons, et vérifier que les types de données sont cohérents avec les variables attendues. Cette étape nous permettra d’identifier les anomalies.

## a) Valeurs manquantes

# Vérification des valeurs manquantes  
# Cette étape permet de détecter la présence de données manquantes  
summary(welfare2018)

## country year hhid grappe   
## Length:7156 Min. :2018 Min. : 1001 Min. : 1.0   
## Class :character 1st Qu.:2018 1st Qu.:151002 1st Qu.:151.0   
## Mode :character Median :2018 Median :300003 Median :300.0   
## Mean :2018 Mean :299934 Mean :299.9   
## 3rd Qu.:2018 3rd Qu.:449010 3rd Qu.:449.0   
## Max. :2018 Max. :598012 Max. :598.0   
##   
## menage vague zae region   
## Min. : 1.000 Min. :1.000 Min. :1.000 Min. : 1.000   
## 1st Qu.: 3.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.: 3.000   
## Median : 6.000 Median :2.000 Median :4.000 Median : 7.000   
## Mean : 6.491 Mean :1.501 Mean :3.466 Mean : 6.781   
## 3rd Qu.: 9.000 3rd Qu.:2.000 3rd Qu.:5.000 3rd Qu.:10.000   
## Max. :12.000 Max. :2.000 Max. :6.000 Max. :14.000   
##   
## milieu hhweight hhsize eqadu1   
## Min. :1.000 Min. : 15.29 Min. : 1.00 Min. : 0.660   
## 1st Qu.:1.000 1st Qu.: 120.57 1st Qu.: 5.00 1st Qu.: 3.960   
## Median :1.000 Median : 203.88 Median : 8.00 Median : 5.982   
## Mean :1.449 Mean : 250.38 Mean : 9.24 Mean : 6.867   
## 3rd Qu.:2.000 3rd Qu.: 322.74 3rd Qu.:12.00 3rd Qu.: 8.720   
## Max. :2.000 Max. :2808.65 Max. :56.00 Max. :41.240   
##   
## eqadu2 hgender hage hmstat   
## Min. : 1.000 Min. :1.000 Min. :17.00 Min. :1.000   
## 1st Qu.: 3.167 1st Qu.:1.000 1st Qu.:41.00 1st Qu.:2.000   
## Median : 4.333 Median :1.000 Median :51.00 Median :2.000   
## Mean : 4.811 Mean :1.262 Mean :51.49 Mean :2.704   
## 3rd Qu.: 5.984 3rd Qu.:2.000 3rd Qu.:62.00 3rd Qu.:3.000   
## Max. :23.706 Max. :2.000 Max. :99.00 Max. :7.000   
## NA's :2   
## hreligion hnation halfab heduc   
## Min. :1.000 Min. : 2.000 Min. :0.0000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.: 7.000 1st Qu.:0.0000 1st Qu.:1.000   
## Median :1.000 Median : 7.000 Median :0.0000 Median :1.000   
## Mean :1.061 Mean : 7.024 Mean :0.4707 Mean :2.188   
## 3rd Qu.:1.000 3rd Qu.: 7.000 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :5.000 Max. :12.000 Max. :1.0000 Max. :9.000   
##   
## hdiploma hhandig hactiv7j hactiv12m   
## Min. : 0.0000 Min. :0.00000 Min. :1.000 Min. :1.00   
## 1st Qu.: 0.0000 1st Qu.:0.00000 1st Qu.:1.000 1st Qu.:1.00   
## Median : 0.0000 Median :0.00000 Median :1.000 Median :1.00   
## Mean : 0.6385 Mean :0.09083 Mean :1.975 Mean :1.49   
## 3rd Qu.: 0.0000 3rd Qu.:0.00000 3rd Qu.:2.000 3rd Qu.:1.00   
## Max. :10.0000 Max. :1.00000 Max. :5.000 Max. :3.00   
##   
## hbranch hsectins hcsp dali   
## Min. : 1.000 Min. :1.000 Min. : 1.000 Min. : 113187   
## 1st Qu.: 1.000 1st Qu.:3.000 1st Qu.: 5.000 1st Qu.: 1143473   
## Median : 6.000 Median :3.000 Median : 9.000 Median : 1746986   
## Mean : 5.324 Mean :2.894 Mean : 7.428 Mean : 2063862   
## 3rd Qu.: 9.000 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 2575472   
## Max. :11.000 Max. :6.000 Max. :10.000 Max. :31295272   
## NA's :1722 NA's :1722 NA's :1722   
## dnal dtot pcexp zzae   
## Min. : 27585 Min. : 278116 Min. : 64161 Min. :296311   
## 1st Qu.: 990977 1st Qu.: 2268859 1st Qu.: 300149 1st Qu.:305745   
## Median : 1632298 Median : 3484846 Median : 442539 Median :326047   
## Mean : 2268143 Mean : 4332005 Mean : 615630 Mean :332244   
## 3rd Qu.: 2771981 3rd Qu.: 5391925 3rd Qu.: 697134 3rd Qu.:348125   
## Max. :221582921 Max. :227607152 Max. :14286279 Max. :391340   
##   
## zref def\_spa def\_temp   
## Min. :333441 Min. :0.8886 Min. :0.9916   
## 1st Qu.:333441 1st Qu.:0.9169 1st Qu.:0.9916   
## Median :333441 Median :0.9778 Median :0.9955   
## Mean :333441 Mean :0.9964 Mean :1.0015   
## 3rd Qu.:333441 3rd Qu.:1.0440 3rd Qu.:1.0089   
## Max. :333441 Max. :1.1736 Max. :1.0147   
##

summary(welfare2021)

## grappe menage country year   
## Min. : 2.0 Min. : 1.000 Length:7120 Min. :2021   
## 1st Qu.:152.0 1st Qu.: 4.000 Class :character 1st Qu.:2021   
## Median :301.0 Median : 7.000 Mode :character Median :2021   
## Mean :300.7 Mean : 7.187 Mean :2021   
## 3rd Qu.:450.0 3rd Qu.:10.000 3rd Qu.:2021   
## Max. :598.0 Max. :19.000 Max. :2021   
##   
## hhid vague month zae   
## Min. : 201 Min. :1.000 Min. :2021-11-01 Min. : 1.000   
## 1st Qu.:15208 1st Qu.:1.000 1st Qu.:2021-12-01 1st Qu.: 5.000   
## Median :30103 Median :2.000 Median :2022-04-01 Median : 7.000   
## Mean :30082 Mean :1.503 Mean :2022-02-16 Mean : 6.712   
## 3rd Qu.:45004 3rd Qu.:2.000 3rd Qu.:2022-05-01 3rd Qu.: 9.000   
## Max. :59812 Max. :2.000 Max. :2022-07-01 Max. :11.000   
##   
## region milieu hhweight hhsize   
## Min. : 1.000 Min. :1.000 Min. : 17.73 Min. : 1.000   
## 1st Qu.: 3.000 1st Qu.:1.000 1st Qu.: 131.07 1st Qu.: 5.000   
## Median : 7.000 Median :1.000 Median : 221.81 Median : 8.000   
## Mean : 6.799 Mean :1.449 Mean : 297.44 Mean : 8.747   
## 3rd Qu.:10.000 3rd Qu.:2.000 3rd Qu.: 380.29 3rd Qu.:11.000   
## Max. :14.000 Max. :2.000 Max. :3081.51 Max. :53.000   
##   
## eqadu1 eqadu2 hgender hage   
## Min. : 0.660 Min. : 1.000 Min. :1.000 Min. : 16.00   
## 1st Qu.: 3.940 1st Qu.: 3.008 1st Qu.:1.000 1st Qu.: 44.00   
## Median : 5.790 Median : 4.180 Median :1.000 Median : 54.00   
## Mean : 6.589 Mean : 4.625 Mean :1.284 Mean : 54.08   
## 3rd Qu.: 8.350 3rd Qu.: 5.762 3rd Qu.:2.000 3rd Qu.: 64.00   
## Max. :40.265 Max. :22.882 Max. :2.000 Max. :101.00   
##   
## hmstat hreligion hnation hethnie   
## Min. :1.000 Min. :1.000 Min. : 4.00 Min. : 1.000   
## 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:13.00 1st Qu.: 1.000   
## Median :2.000 Median :1.000 Median :13.00 Median : 3.000   
## Mean :2.805 Mean :1.058 Mean :12.95 Mean : 3.096   
## 3rd Qu.:3.000 3rd Qu.:1.000 3rd Qu.:13.00 3rd Qu.: 3.000   
## Max. :7.000 Max. :5.000 Max. :18.00 Max. :13.000   
## NA's :82   
## halfa halfa2 heduc hdiploma   
## Min. :0.0000 Min. :0.0000 Min. :1.000 Min. : 0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:1.000 1st Qu.: 0.0000   
## Median :1.0000 Median :0.0000 Median :1.000 Median : 0.0000   
## Mean :0.5117 Mean :0.4947 Mean :2.112 Mean : 0.5622   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:3.000 3rd Qu.: 0.0000   
## Max. :1.0000 Max. :1.0000 Max. :9.000 Max. :10.0000   
##   
## hhandig hactiv7j hactiv12m hbranch   
## Min. :0.00000 Min. :1.000 Min. :1.000 Min. : 1.00   
## 1st Qu.:0.00000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.: 1.00   
## Median :0.00000 Median :1.000 Median :1.000 Median : 5.00   
## Mean :0.08919 Mean :2.067 Mean :1.382 Mean : 6.36   
## 3rd Qu.:0.00000 3rd Qu.:5.000 3rd Qu.:1.000 3rd Qu.: 8.00   
## Max. :1.00000 Max. :5.000 Max. :3.000 Max. :930.00   
## NA's :1838   
## hsectins hcsp dali dnal   
## Min. :1.000 Min. : 1.000 Min. : 64205 Min. : 129749   
## 1st Qu.:3.000 1st Qu.: 9.000 1st Qu.: 1325492 1st Qu.: 972937   
## Median :3.000 Median : 9.000 Median : 1956762 Median : 1595027   
## Mean :2.937 Mean : 7.717 Mean : 2276262 Mean : 2029004   
## 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 2840293 3rd Qu.: 2556010   
## Max. :6.000 Max. :10.000 Max. :15144512 Max. :19917053   
## NA's :1359 NA's :1326   
## dtot pcexp zzae zref   
## Min. : 235210 Min. : 57610 Min. :331925 Min. :369666   
## 1st Qu.: 2432066 1st Qu.: 328408 1st Qu.:335734 1st Qu.:369666   
## Median : 3614746 Median : 472561 Median :373110 Median :369666   
## Mean : 4305266 Mean : 621198 Mean :371487 Mean :369666   
## 3rd Qu.: 5359042 3rd Qu.: 724571 3rd Qu.:385554 3rd Qu.:369666   
## Max. :29248050 Max. :9990532 Max. :424179 Max. :369666   
##   
## def\_spa def\_temp def\_temp\_prix2021m11 def\_temp\_cpi   
## Min. :0.8979 Min. :0.9455 Min. :1.000 Min. :0.9743   
## 1st Qu.:0.9082 1st Qu.:0.9833 1st Qu.:1.001 1st Qu.:0.9752   
## Median :1.0093 Median :0.9949 Median :1.020 Median :0.9935   
## Mean :1.0049 Mean :0.9981 Mean :1.020 Mean :0.9935   
## 3rd Qu.:1.0430 3rd Qu.:1.0200 3rd Qu.:1.027 3rd Qu.:1.0009   
## Max. :1.1475 Max. :1.0590 Max. :1.087 Max. :1.0587   
##   
## def\_temp\_adj zali0 dtet monthly\_cpi   
## Min. :0.9606 Min. :196233 Min. : 50712 Min. :117.2   
## 1st Qu.:0.9991 1st Qu.:196233 1st Qu.: 317994 1st Qu.:117.4   
## Median :1.0109 Median :196233 Median : 469028 Median :119.6   
## Mean :1.0140 Mean :196233 Mean : 641538 Mean :119.6   
## 3rd Qu.:1.0363 3rd Qu.:196233 3rd Qu.: 746041 3rd Qu.:120.4   
## Max. :1.0760 Max. :196233 Max. :10366096 Max. :127.4   
##   
## cpi2017 icp2017 dollars   
## Min. :1.097 Min. :238.6 Min. : 0.5376   
## 1st Qu.:1.097 1st Qu.:238.6 1st Qu.: 3.2473   
## Median :1.097 Median :238.6 Median : 4.7915   
## Mean :1.097 Mean :238.6 Mean : 6.5526   
## 3rd Qu.:1.097 3rd Qu.:238.6 3rd Qu.: 7.6134   
## Max. :1.097 Max. :238.6 Max. :102.5809   
##

# Calcul du pourcentage de valeurs manquantes pour chaque colonne  
# Nous calculons ici le pourcentage de valeurs manquantes dans chaque colonne des deux base. Si les valeurs manquantes sont faibles (moins de 5-10%), une imputation (par la médiane ou la moyenne) pourrait être une solution. Sinon, il peut être nécessaire de supprimer les colonnes ou de les traiter autrement  
  
colSums(is.na(welfare2018)) / nrow(welfare2018) \* 100

## country year hhid grappe menage vague   
## 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000   
## zae region milieu hhweight hhsize eqadu1   
## 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000   
## eqadu2 hgender hage hmstat hreligion hnation   
## 0.00000000 0.00000000 0.00000000 0.02794857 0.00000000 0.00000000   
## halfab heduc hdiploma hhandig hactiv7j hactiv12m   
## 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000   
## hbranch hsectins hcsp dali dnal dtot   
## 24.06372275 24.06372275 24.06372275 0.00000000 0.00000000 0.00000000   
## pcexp zzae zref def\_spa def\_temp   
## 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000

colSums(is.na(welfare2021)) / nrow(welfare2021) \* 100

## grappe menage country   
## 0.000000 0.000000 0.000000   
## year hhid vague   
## 0.000000 0.000000 0.000000   
## month zae region   
## 0.000000 0.000000 0.000000   
## milieu hhweight hhsize   
## 0.000000 0.000000 0.000000   
## eqadu1 eqadu2 hgender   
## 0.000000 0.000000 0.000000   
## hage hmstat hreligion   
## 0.000000 0.000000 0.000000   
## hnation hethnie halfa   
## 0.000000 1.151685 0.000000   
## halfa2 heduc hdiploma   
## 0.000000 0.000000 0.000000   
## hhandig hactiv7j hactiv12m   
## 0.000000 0.000000 0.000000   
## hbranch hsectins hcsp   
## 25.814607 19.087079 18.623596   
## dali dnal dtot   
## 0.000000 0.000000 0.000000   
## pcexp zzae zref   
## 0.000000 0.000000 0.000000   
## def\_spa def\_temp def\_temp\_prix2021m11   
## 0.000000 0.000000 0.000000   
## def\_temp\_cpi def\_temp\_adj zali0   
## 0.000000 0.000000 0.000000   
## dtet monthly\_cpi cpi2017   
## 0.000000 0.000000 0.000000   
## icp2017 dollars   
## 0.000000 0.000000

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

## b) Doublons

# Vérification des doublons  
# On vérifie s'il y a des doublons dans les deux bases  
sum(duplicated(welfare2018)) # Doublons dans la base 2018

## [1] 0

sum(duplicated(welfare2021)) # Doublons dans la base 2021

## [1] 0

Aucun doublon n’a été détecté dans les deux bases

# 5. Analyse comparative des bases des variables et modalités des deux bases welfare

Ayant visualisé les données et vérifié la présence d’anomalies, nous allons analyser les similitudes et les différences entre les bases welfare 2018 et 2021. Cette étape nous permettra de détecter des différences d’annotation entre les variables et/ou modalités présentes entre les deux bases.

variables\_2018 = colnames(welfare2018) #Liste des variables de la base welfare 2018  
  
variables\_2021 = colnames(welfare2021) #Liste des variables de la base welfare 2021  
  
# Identifions les variables communes  
variables\_communes <- intersect(variables\_2018, variables\_2021) # Variables communes aux deux bases  
  
# Variables présentes exclusivement en 2018  
vars\_exclu2018 <- setdiff(variables\_2018, variables\_2021) #Pour garder les variables de 2018 qui ne sont pas en 2021  
  
# Variables présentes exclusivement en 2021  
vars\_exclu2021 <- setdiff(variables\_2021, variables\_2018) #Pour garder les variables de 2021 qui ne sont pas en 2018

#Visualisation des variables propres à chaque base  
print(vars\_exclu2018)

## [1] "halfab"

print(vars\_exclu2021)

## [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 fait apparaitre que les deux bases ont 34 variables en commun, une seule variable est unique pour base welfare 2018, soit halfab et 13 pour celle de 2021 ( “month”, “hethnie”, “halfa”, “halfa2”, “def\_temp\_prix2021m11”, “def\_temp\_cpi”, “def\_temp\_adj”, “zali0”, “dtet”, “monthly\_cpi”, “cpi2017”, icp2017”, “dollars” )

On remarque pour la variable alphabétisation, il s’agit juste d’une différence d’orthographe. Procédons à la renomination de la variable de 2021 en halfab et intégrons là aux variables communes

# Changement halfa vers halfab  
  
colnames(welfare2021)[colnames(welfare2021)=="halfa"] <- "halfab"  
  
# Ajout à la liste des variables communes  
  
variables\_communes <- append(variables\_communes,"halfab")

Maintenant, étant donné que les variables spécifiques à une seule base sont isolées et que les variables communes portent les mêmes noms dans les deux bases, nous allons vérifier la concordance de leur labellisation.

## a) Recensement des incohérences

discord\_label\_var <- c()   
  
for (variable in variables\_communes) { #parcourir les variables en communs dans les deux bases  
   
 if(labelled::is.labelled(welfare2018[[variable]])){ #vérifier si la variable en 2018 est labellisée  
   
 label\_val\_2018 <- labelled::val\_labels(welfare2018[[variable]]) #recupérer les labels de la variable en 2018  
  
 }else{  
 label\_val\_2018 <- NULL #Mettre vide dans le cas ou la variable en 2018 n'est pas labellisée  
 }  
   
 if(labelled::is.labelled(welfare2021[[variable]])){ #vérifier si la variable en 2021 est labellisée  
   
 label\_val\_2021 <- labelled::val\_labels(welfare2021[[variable]]) #recupérer les labels de la variable en 2021  
   
 }else{  
   
 label\_val\_2021 <- NULL #Mettre vide dans le cas ou la variable en 2021 n'est pas labellisée  
 }  
   
 if(!identical(label\_val\_2018, label\_val\_2021)){ #Vérifier si les labels de la variable sont identiques entre 2018 et 2021  
   
 discord\_label\_var <- append(discord\_label\_var,variable) #Si les labels diffèrent, alors ajouter le nom de la variable à la liste créer  
 print(variable)  
   
 }  
}

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

Nous remarquons que six(6) variables ne sont pas labellisées de la même manière dans les deux bases welfare 2018 et 2021, soit “zae”, “hnation”, “hdiploma”, “hactiv7j”, “hbranch” et “hcsp”. Pour corriger ces incohérences, nous allons traiter une à une ces variables, les recoder et aligner leurs modalités selon les incohérences détextées.

## b) Traitement des incohérences

### i) Traitement de la variable **hnation**

#### Visualisation de la distribution

#voir la distribution de hnation en 2018  
labelled::val\_labels(welfare2018$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

#Voir la distribution de hnation en 2021  
labelled::val\_labels(welfare2021$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

welfare2021 %>%  
 to\_factor() %>% #labéliser  
 select(hnation) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

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

#### Detection de l’incoherence

Certains pays qui étaient inclus dans autres cedeao en 2018 sont devenus des modalités comme la Guinée

#### Correction de l’incohérence

Pour ce faire nous allons recoder la variable dans welfare 2021 et combiner les modalités qui étaient dans autre cedeao en 2018

welfare2021 <- welfare2021 %>%  
 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  
 ))

Maintenant que le recodage est effectué on peut aisément affecter les labels de 2018 à ceux de 2021.

labelled::val\_labels(welfare2021$hnation) <- labelled::val\_labels(welfare2018$hnation)

#### Vérification des changements

welfare2021 %>%  
 to\_factor() %>% #labéliser  
 select(hnation) %>% #selection des variables qui vont s'afficher  
 tbl\_summary(missing = "always", #afficher les valeurs manquantes  
 missing\_text = "Valeurs manquantes")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Nationalite du CM |  |
| Benin | 0 (0%) |
| Burkina Faso | 0 (0%) |
| Côte d'Ivoire | 1 (<0.1%) |
| Guinée Bissau | 8 (0.1%) |
| Mali | 18 (0.3%) |
| Niger | 2 (<0.1%) |
| Sénégal | 7,038 (99%) |
| Togo | 1 (<0.1%) |
| Nigéria | 1 (<0.1%) |
| Autre CEDEAO | 42 (0.6%) |
| Autre Afrique | 7 (<0.1%) |
| Autre pays hors Afrique | 2 (<0.1%) |
| Valeurs manquantes | 0 |
| *1*n (%) | |

### ii) Traitement de la variable **hdiploma**

#### Visualisation de la distribution

labelled::val\_labels(welfare2018$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

labelled::val\_labels(welfare2021$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

welfare2018 %>%   
 to\_factor() %>%   
 select(hdiploma) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,156***1* |
| --- | --- |
| Diplome du CM |  |
| Aucun | 5,697 (80%) |
| CEP/CFEE | 587 (8.2%) |
| BEPC/BFEM | 359 (5.0%) |
| cap | 52 (0.7%) |
| bt | 16 (0.2%) |
| bac | 154 (2.2%) |
| DEUG, DUT, BTS | 49 (0.7%) |
| Licence | 83 (1.2%) |
| Maitrise | 63 (0.9%) |
| Master/DEA/DESS | 65 (0.9%) |
| Doctorat/Phd | 31 (0.4%) |
| NA | 0 |
| *1*n (%) | |

#### Detection de l’incoherence

Il s’agit juste de changements de labels

#### Correction de l’incohérence

Pour cela il suffira d’affecter les labels de hdiploma dans welfare 2021 à ceux de 2018

val\_labels(welfare2018$hdiploma) <- val\_labels(welfare2021$hdiploma)

#### Vérification des changements

welfare2018 %>%   
 to\_factor() %>%   
 select(hdiploma) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,156***1* |
| --- | --- |
| Diplome du CM |  |
| Aucun | 5,697 (80%) |
| cepe | 587 (8.2%) |
| bepc | 359 (5.0%) |
| cap | 52 (0.7%) |
| bt | 16 (0.2%) |
| bac | 154 (2.2%) |
| DEUG, DUT, BTS | 49 (0.7%) |
| Licence | 83 (1.2%) |
| Maitrise | 63 (0.9%) |
| Master/DEA/DESS | 65 (0.9%) |
| Doctorat/Phd | 31 (0.4%) |
| NA | 0 |
| *1*n (%) | |

### iii) Traitement de la variable **hactiv7j**

#### Visualisation de la distribution

labelled::val\_labels(welfare2018$hactiv7j)

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

labelled::val\_labels(welfare2021$hactiv7j)

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

welfare2018 %>%   
 to\_factor() %>%   
 select(hactiv7j) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,156***1* |
| --- | --- |
| Activite 7 jours du CM |  |
| Occupe | 5,362 (75%) |
| Chomeur | 44 (0.6%) |
| TF cherchant emploi | 3 (<0.1%) |
| TF cherchant pas | 60 (0.8%) |
| Inactif | 1,687 (24%) |
| Moins de 5 ans | 0 (0%) |
| NA | 0 |
| *1*n (%) | |

#### Detection de l’incoherence

Les labels sont les mêmes mais l’ordre des modalités n’est pas le même, il y a don erreur dans le codage.

#### Correction de l’incohérence

Pour cela nous allons changer l’ordre des modalités de sorte à ce qu’on puisse retrouver les mêmes dans welfare 2021. Ainsi l’affectation des labels de 2021 à 2018 sera possible.

welfare2018 <- welfare2018 %>%  
 mutate(hactiv7j = dplyr::recode(hactiv7j,  
 `2` = 4,   
 `3` = 2,   
 `4` = 3))

#Affectation des labels de hactiv7j 2021 à ceux de 2018  
val\_labels(welfare2018$hactiv7j) <- val\_labels(welfare2021$hactiv7j)

#### Vérification des changements

welfare2018 %>%   
 to\_factor() %>%   
 select(hactiv7j) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,156***1* |
| --- | --- |
| Activite 7 jours du CM |  |
| Occupe | 5,362 (75%) |
| TF cherchant emploi | 3 (<0.1%) |
| TF cherchant pas | 60 (0.8%) |
| Chomeur | 44 (0.6%) |
| Inactif | 1,687 (24%) |
| Moins de 5 ans | 0 (0%) |
| NA | 0 |
| *1*n (%) | |

### iv) Traitement de la variable **hbranch**

#### Visualisation de la distribution

labelled::val\_labels(welfare2018$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

labelled::val\_labels(welfare2021$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

welfare2018 %>%   
 to\_factor() %>%   
 select(hbranch) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,156***1* |
| --- | --- |
| Branche activite 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 (%) | |

#### Detection de l’incoherence

On remarque qu’en 2021 le secteur de la sylvopasture a été rajouté et combiné à l’élevage et la pêche, ce qui est différent pour 2018.

#### Correction de l’incohérence

Cette erreur peut être corrigée en faisant juste une affectation comme pour les autres variables.

val\_labels(welfare2018$hbranch) <- val\_labels(welfare2021$hbranch)

#### Vérification des changements

welfare2018 %>%   
 to\_factor() %>%   
 select(hbranch) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,156***1* |
| --- | --- |
| Branche activite 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 (%) | |

### v) Traitement de la variable **hcsp**

#### Visualisation de la distribution

labelled::val\_labels(welfare2018$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

labelled::val\_labels(welfare2021$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

#### Detection de l’incoherence

La différence se situe juste dans la labellisation de la neuvième modalité.

#### Correction de l’incohérence

Ici nous affecterons les labels de 2018 à ceux de 2021

#Affectation des labels de 2018 à 2021  
val\_labels(welfare2021$hcsp) <- val\_labels(welfare2018$hcsp)

#### Vérification des changements

welfare2021 %>%   
 to\_factor() %>%   
 select(hcsp) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| CSP 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 (%) | |

### vi) Traitement de la variable **zae**

#### Visualisation de la distribution

labelled::val\_labels(welfare2018$zae)

## NULL

labelled::val\_labels(welfare2021$zae)

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

welfare2018 %>%   
 to\_factor() %>%   
 select(zae) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,156***1* |
| --- | --- |
| Zone agroecologique |  |
| 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 (%) | |

welfare2021 %>%   
 select(zae) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

## ! Column(s) "zae" are class "haven\_labelled".  
## ℹ This is an intermediate datastructure 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>

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Zone agroecologique |  |
| 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 (%) | |

welfare2021 %>%   
 to\_factor() %>%   
 select(zae) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

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

#### Detection de l’incoherence

Là nous remarquons que pour la base welfare 2018, la variable zae n’est non seulement pas labellisée, mais les codes diffèrent pour les deux bases, soit 1, 2, 3 , 4 , 5, et 6 pour 2018 et 1, 3, 5, 7, 9 et 11 pour 2021. En plus Kedougou et Dakar sont inversés en 2018, en faisant attention àla fréquence dans les deux bases.

#### Correction de l’incohérence

Nous allons d’abord faire correspondre les codes de 2018 à ceux de 2021 avant de pouvoir passer à l’affectation pour labelliser la variable dans welfare 2018 comme pour elfare 2021.

#####Retablissons l’ordre pour Kedougou et Dakar en 2018

welfare2018 <- welfare2018 %>%  
 mutate(zae = dplyr::recode(zae,  
 `1` = 6, # Kédougou remplace Dakar  
 `2` = 2, # Saint-Louis-Matam  
 `3` = 3,# Thies-Diourbel-Louga  
 `4` = 4, # Kaolack-Fatick-Kaffrine  
 `5` = 5, # Ziguinchor-Tamba-Kolda-Sédhiou  
 `6` = 1 # Dakar devient Kedougou  
 ))

##### Verfifions le changement en 2018 apres mutation Kedougou et Dakar

welfare2018 %>%   
 to\_factor() %>%   
 select(zae) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

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

# Harmoniser les codes impairs en 1:6 pour 2021  
welfare2021 <- welfare2021 %>%  
 mutate(zae = dplyr::recode(zae,  
 `1` = 1,  
 `3` = 2,  
 `5` = 3,  
 `7` = 4,  
 `9` = 5,  
 `11` = 6  
 ))

#### Vérification

#pour savoir si le recodage a marché  
welfare2021 %>%   
 select(zae) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

## ! Column(s) "zae" are class "haven\_labelled".  
## ℹ This is an intermediate datastructure 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>

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Zone agroecologique |  |
| 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 (%) | |

welfare2021 <- welfare2021 %>%  
 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"  
 ))

welfare2018 <- welfare2018 %>%  
 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

#Affiche la distribution en 2018 après l'harmonisation finale  
welfare2018 %>%   
 to\_factor() %>%   
 select(zae) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

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

#Affiche la distribution en 2021 après l'harmonisation finale  
welfare2021 %>%   
 to\_factor() %>%   
 select(zae) %>%   
 tbl\_summary(missing = "always",  
 missing\_text = "NA")

| **Characteristic** | **N = 7,120***1* |
| --- | --- |
| Zone agroecologique |  |
| 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 (%) | |

Le codage et la labellisation de la variable zae sont maintenant harmonisées dans les deux bases. Maintenant Vérifions qu’il n’existe plus aucune différence dans la labellisation ou le codage entre les variables communes des deux bases

# 6. Vérification du traitement

discord\_label\_var <- c()   
  
for (variable in variables\_communes) { #parcourir les variables en communs dans les deux bases  
   
 if(labelled::is.labelled(welfare2018[[variable]])){ #vérifier si la variable en 2018 est labellisée  
   
 label\_val\_2018 <- labelled::val\_labels(welfare2018[[variable]]) #recupérer les labels de la variable en 2018  
  
 }else{  
 label\_val\_2018 <- NULL #Mettre vide dans le cas ou la variable en 2018 n'est pas labellisée  
 }  
   
 if(labelled::is.labelled(welfare2021[[variable]])){ #vérifier si la variable en 2021 est labellisée  
   
 label\_val\_2021 <- labelled::val\_labels(welfare2021[[variable]]) #recupérer les labels de la variable en 2021  
   
 }else{  
   
 label\_val\_2021 <- NULL #Mettre vide dans le cas ou la variable en 2021 n'est pas labellisée  
 }  
   
 if(!identical(label\_val\_2018, label\_val\_2021)){ #Vérifier si les labels de la variable sont identiques entre 2018 et 2021  
   
 discord\_label\_var <- append(discord\_label\_var,variable) #Si les labels diffèrent, alors ajouter le nom de la variable à la liste créer  
 print(variable)  
   
 }  
}

On a pu vérifier qu’il n’existe plus aucun incohérence entre les deux bases. Le traitement est donc achevé et il est possible de passer au merge des deux bases

# 7. Merge des bases traitées welfare 2018 et welfare 2021

#Merge des bases  
library(dplyr)  
welfare\_final <- bind\_rows(welfare2018, welfare2021)

Maintenant nous allons visualiser la structure de la base mergée, afficher ses variables et ses dix premières lignes

colnames(welfare\_final) #Liste des variables

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

head(welfare\_final, 10) # Affiche les 10 premières lignes

## # A tibble: 10 × 47  
## country year hhid grappe menage vague zae region milieu hhweight hhsize  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl+l> <dbl+l> <dbl> <dbl>  
## 1 SEN 2018 1001 1 1 1 Dakar 1 [dak… 1 [Urb… 1750. 2  
## 2 SEN 2018 1002 1 2 1 Dakar 1 [dak… 1 [Urb… 1750. 2  
## 3 SEN 2018 1003 1 3 1 Dakar 1 [dak… 1 [Urb… 1750. 1  
## 4 SEN 2018 2001 2 1 2 Dakar 1 [dak… 1 [Urb… 266. 10  
## 5 SEN 2018 2002 2 2 2 Dakar 1 [dak… 1 [Urb… 266. 6  
## 6 SEN 2018 2003 2 3 2 Dakar 1 [dak… 1 [Urb… 266. 4  
## 7 SEN 2018 2004 2 4 2 Dakar 1 [dak… 1 [Urb… 266. 3  
## 8 SEN 2018 2005 2 5 2 Dakar 1 [dak… 1 [Urb… 266. 2  
## 9 SEN 2018 2006 2 6 2 Dakar 1 [dak… 1 [Urb… 266. 3  
## 10 SEN 2018 2007 2 7 2 Dakar 1 [dak… 1 [Urb… 266. 1  
## # ℹ 36 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>,  
## # month <date>, hethnie <dbl+lbl>, halfa2 <dbl+lbl>, …

str(welfare\_final) #Structure de la base mergée

## tibble [14,276 × 47] (S3: tbl\_df/tbl/data.frame)  
## $ country : chr [1:14276] "SEN" "SEN" "SEN" "SEN" ...  
## $ year : num [1:14276] 2018 2018 2018 2018 2018 ...  
## $ hhid : num [1:14276] 1001 1002 1003 2001 2002 ...  
## $ grappe : num [1:14276] 1 1 1 2 2 2 2 2 2 2 ...  
## $ menage : num [1:14276] 1 2 3 1 2 3 4 5 6 7 ...  
## $ vague : num [1:14276] 1 1 1 2 2 2 2 2 2 2 ...  
## $ zae : chr [1:14276] "Dakar" "Dakar" "Dakar" "Dakar" ...  
## $ region : dbl+lbl [1:14276] 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 [1:14276] 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 [1:14276] 1750 1750 1750 266 266 ...  
## $ hhsize : num [1:14276] 2 2 1 10 6 4 3 2 3 1 ...  
## $ eqadu1 : num [1:14276] 1.32 1.45 1 7.66 4.28 ...  
## $ eqadu2 : num [1:14276] 1.61 1.61 1 5.54 3.25 ...  
## $ hgender : dbl+lbl [1:14276] 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 [1:14276] 59 59 27 85 40 68 31 54 67 42 ...  
## $ hmstat : dbl+lbl [1:14276] 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 [1:14276] 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 [1:14276] 7, 12, 7, 7, 7, 7, 11, 7, 7, 7, 7, 7, 7, ...  
## ..@ 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" ...  
## ..@ label : chr "Nationalite du CM"  
## $ halfab : dbl+lbl [1:14276] 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0...  
## ..@ labels: Named num [1:2] 0 1  
## .. ..- attr(\*, "names")= chr [1:2] "Non" "Oui"  
## ..@ label : chr "Alphabetisation du CM"  
## $ heduc : dbl+lbl [1:14276] 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 [1:14276] 1, 6, 10, 0, 0, 0, 10, 0, 0, 1, 0, 0, 0, ...  
## ..@ labels: Named num [1:11] 0 1 2 3 4 5 6 7 8 9 ...  
## .. ..- attr(\*, "names")= chr [1:11] "Aucun" "cepe" "bepc" "cap" ...  
## ..@ label : chr "Diplome du CM"  
## $ hhandig : dbl+lbl [1:14276] 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 [1:14276] 1, 1, 1, 5, 1, 5, 1, 1, 5, 1, 1, 1, 1, 1, 1, 1, 1, 1...  
## ..@ labels: Named num [1:6] 1 2 3 4 5 6  
## .. ..- attr(\*, "names")= chr [1:6] "Occupe" "TF cherchant emploi" "TF cherchant pas" "Chomeur" ...  
## ..@ label : chr "Activite 7 jours du CM"  
## $ hactiv12m : dbl+lbl [1:14276] 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 [1:14276] 6, 6, 11, NA, 6, NA, 9, 11, NA, 6, 7, 8, 6, ...  
## ..@ 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." ...  
## ..@ label : chr "Branche activite du CM"  
## $ hsectins : dbl+lbl [1:14276] 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 [1:14276] 4, 1, 7, NA, 4, NA, 6, 5, NA, 9, 3, 4, 9, ...  
## ..@ 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" ...  
## ..@ label : chr "CSP du CM"  
## $ dali : num [1:14276] 1978488 1886313 387989 6568518 2362250 ...  
## $ dnal : num [1:14276] 13139233 8894985 5055663 4889539 3930703 ...  
## $ dtot : num [1:14276] 15117721 10781298 5443652 11458057 6292954 ...  
## $ pcexp : num [1:14276] 6440507 4593088 4638249 976281 893649 ...  
## $ zzae : num [1:14276] 391340 391340 391340 391340 391340 ...  
## $ zref : num [1:14276] 333441 333441 333441 333441 333441 ...  
## $ def\_spa : num [1:14276] 1.17 1.17 1.17 1.17 1.17 ...  
## $ def\_temp : num [1:14276] 1.015 1.012 1.015 0.992 0.992 ...  
## $ month : Date[1:14276], format: NA NA ...  
## $ hethnie : dbl+lbl [1:14276] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...  
## ..@ 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" ...  
## $ halfa2 : dbl+lbl [1:14276] NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...  
## ..@ 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"  
## $ def\_temp\_prix2021m11: num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ def\_temp\_cpi : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ def\_temp\_adj : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ zali0 : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ dtet : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ monthly\_cpi : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ cpi2017 : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ icp2017 : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...  
## $ dollars : num [1:14276] NA NA NA NA NA NA NA NA NA NA ...

write\_dta(welfare\_final, "../Données/welfare\_final.dta") #Enregistrement local

# Trier par identifiant ménage + année  
welfare\_final <- welfare\_final %>%  
 arrange(hhid, year)

head(welfare\_final, 10)

## # A tibble: 10 × 47  
## country year hhid grappe menage vague zae region milieu hhweight hhsize  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <dbl+l> <dbl+l> <dbl> <dbl>  
## 1 SEN 2021 201 2 1 2 Dakar 1 [dak… 1 [Urb… 290. 9  
## 2 SEN 2021 203 2 3 2 Dakar 1 [dak… 1 [Urb… 386. 3  
## 3 SEN 2021 204 2 4 2 Dakar 1 [dak… 1 [Urb… 290. 3  
## 4 SEN 2021 205 2 5 2 Dakar 1 [dak… 1 [Urb… 386. 3  
## 5 SEN 2021 206 2 6 2 Dakar 1 [dak… 1 [Urb… 290. 3  
## 6 SEN 2021 207 2 7 2 Dakar 1 [dak… 1 [Urb… 290. 1  
## 7 SEN 2021 208 2 8 2 Dakar 1 [dak… 1 [Urb… 290. 1  
## 8 SEN 2021 212 2 12 2 Dakar 1 [dak… 1 [Urb… 290. 6  
## 9 SEN 2021 213 2 13 2 Dakar 1 [dak… 1 [Urb… 386. 5  
## 10 SEN 2021 214 2 14 2 Dakar 1 [dak… 1 [Urb… 386. 4  
## # ℹ 36 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>,  
## # month <date>, hethnie <dbl+lbl>, halfa2 <dbl+lbl>, …

# 8. Statistiques descriptives sur la base welfare :

Maintenant que nous avons une vue sur la base mergée, nous allons sortir quelques statistiques descriptives en utilisant cette base. Nous allons faire une analyse univariée sur la dépense par tête (en cfa) et une analyse bivariée mettant en rapport la dépense par tête (en fcfa) et la région, pour avoir un aperçu sur le niveau de vie selon la région.

## a) Analyse univariée

Montrons les caractéristiques de la variable **pcexp** ou dépense par tête en FCFA

welfare\_final %>%  
 select(pcexp) %>%  
 tbl\_summary(  
 statistic = list(all\_continuous() ~ "{mean} ({sd}) [min: {min}; max: {max}]"),  
 digits = all\_continuous() ~ 0,  
 missing = "no"  
 ) %>%  
 modify\_header(label ~ "Dépense par tête (FCFA)")

| Dépense par tête (FCFA) | **N = 14,276***1* |
| --- | --- |
| pcexp | 618,407 (582,644) [min: 57,610; max: 14,286,279] |
| *1*Mean (SD) [min: Min; max: Max] | |

On constate que la dépense moyenne par tête dans l’échantillon est estimée à 618 407 FCFA, avec un écart-type de 582 644 FCFA, ce qui indique une forte hétérogénéité des niveaux de vie entre les individus. Le minimum observé est de 57 610 FCFA, tandis que le maximum atteint 14 286 279 FCFA, révélant une distribution très étalée et potentiellement asymétrique, marquée par la présence de ménages très dépensiers. Cette dispersion suggère des inégalités importantes en termes de consommation individuelle

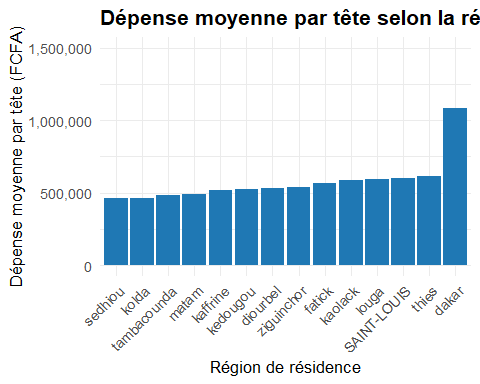
## b) Analyse bivariée

Ici nous allons analyser la distribution de la dépense par tête suivant la région.

library(gtsummary)  
library(dplyr)  
library(haven) #Pour que les labels apparaissent  
  
welfare\_final %>%  
 mutate(region = haven::as\_factor(region)) %>%  
 select(region, pcexp) %>%  
 tbl\_summary(  
 by = region,  
 statistic = list(all\_continuous() ~ "{mean} ({sd}) [min: {min}; max: {max}]"),  
 digits = all\_continuous() ~ 0,  
 missing = "no"  
 ) %>%  
 add\_overall() %>%  
 modify\_header(label ~ "Région de résidence")

| Région de résidence | **Overall** N = 14,276*1* | **dakar** N = 2,025*1* | **ziguinchor** N = 948*1* | **diourbel** N = 1,104*1* | **SAINT-LOUIS** N = 1,007*1* | **tambacounda** N = 864*1* | **kaolack** N = 1,054*1* | **thies** N = 1,137*1* | **louga** N = 960*1* | **fatick** N = 910*1* | **kolda** N = 864*1* | **matam** N = 816*1* | **kaffrine** N = 863*1* | **kedougou** N = 908*1* | **sedhiou** N = 816*1* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| pcexp | 618,407 (582,644) [min: 57,610; max: 14,286,279] | 1,084,479 (919,563) [min: 162,244; max: 11,601,709] | 536,783 (434,480) [min: 57,610; max: 5,271,182] | 535,110 (347,354) [min: 119,643; max: 4,694,128] | 598,863 (477,865) [min: 83,650; max: 7,089,593] | 484,043 (605,657) [min: 71,963; max: 14,286,279] | 588,986 (478,862) [min: 64,161; max: 6,280,066] | 616,116 (481,343) [min: 63,897; max: 8,204,014] | 594,655 (447,856) [min: 126,487; max: 4,145,988] | 563,862 (441,943) [min: 92,848; max: 5,044,186] | 466,439 (369,203) [min: 76,139; max: 3,811,420] | 487,646 (340,393) [min: 79,123; max: 3,216,407] | 520,167 (574,650) [min: 86,234; max: 9,990,532] | 526,022 (478,516) [min: 62,654; max: 3,665,106] | 464,043 (434,659) [min: 90,050; max: 5,765,535] |
| *1*Mean (SD) [min: Min; max: Max] | | | | | | | | | | | | | | | |

library(ggplot2)  
library(dplyr)  
library(haven)  
  
welfare\_final <- welfare\_final %>%  
 mutate(region = haven::as\_factor(region))  
  
# Calculer la moyenne par région  
depense\_moyenne <- welfare\_final %>%  
 group\_by(region) %>%  
 summarise(pcexp\_moy = mean(pcexp, na.rm = TRUE)) %>%  
 arrange(desc(pcexp\_moy))  
  
# Créer le barplot  
ggplot(depense\_moyenne, aes(x = reorder(region, pcexp\_moy), y = pcexp\_moy)) +  
 geom\_col(fill = "#1f78b4") +  
 labs(  
 title = "Dépense moyenne par tête selon la région",  
 x = "Région de résidence",  
 y = "Dépense moyenne par tête (FCFA)"  
 ) +  
 scale\_y\_continuous(labels = scales::comma, limits = c(0, 1500000)) + # Ajuste le max en ordonnées si besoin  
 theme\_minimal(base\_size = 13) +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 plot.title = element\_text(face = "bold")  
 )



L’analyse de la dépense par tête selon la région révèle des disparités notables entre les zones géographiques. En moyenne, la dépense par tête dans l’échantillon est estimée à 618 407 FCFA, mais cette moyenne cache de fortes variations régionales. La région de Dakar affiche de loin la dépense la plus élevée, avec une moyenne de 1 084 479 FCFA, soit presque le double de la moyenne nationale, ce qui reflète son statut de capitale économique. À l’opposé, des régions comme Matam (466 439 FCFA) et Sédhiou (464 043 FCFA) présentent les niveaux de dépense par tête les plus faibles. Des régions telles que Thiès, Louga ou Fatick se situent autour de la moyenne nationale, tandis que d’autres comme Kolda ou Kaffrine sont légèrement en retrait. Ces écarts illustrent une inégalité spatiale du niveau de vie, suggérant que les conditions économiques varient significativement d’un territoire à l’autre, appelant à une attention particulière en matière de politiques publiques de rééquilibrage régional.