RÉPUBLIQUE DU SÉNÉGAL

Un Peuple - Un But - Une Foi

MINISTÈRE DE L'ÉCONOMIE
DU PLAN ET DE LA COOPÉRATION

ANSD

Agence nationale de la Statistique et de la Démographie (ANSD)

ensae

Ecole nationale de la Statistique et de l'Analyse économique Pierre Ndiaye (ENSAE)

Semestre 2 :Projet statistique avec R

TP 5 : MERGING

Rédigé par :
Alioune Abdou Salam Kane
Khadidiatou Diakhaté
Ange Emilson Rayan Rahérinasolo
Awa Diaw
Elèves en ISE 1 cycle long

Sous la supervision de :
M. Aboubacre HEMA
Research Analyst

Année scolaire : 2024/2025

Sommaire

Introduction

Le projet **TP4 : MERGING** a pour objectif principal de préparer un fichier enrichi en fusionnant les données provenant de deux sources importantes : l'**Enquête Harmonisée sur les Conditions de Vie des Ménages (EHCVM) 2021** et le **Humanitarian Data Exchange (HDX)**. Le défi majeur de ce projet réside dans la gestion des différences de format et de nomenclature des communes entre ces deux bases de données, ce qui nécessite un processus minutieux de nettoyage et de préparation des données.

Afin de mener à bien cette tâche, nous avons utilisé **Excel** pour effectuer les premières étapes de traitement des données. Tout d'abord, nous avons créé un troisième fichier vecteur à deux variables où chacune servait de clé unique pour la fucion avec les 2 datasets initiaux.

Cette approche nous a évité d'altérer les variables de base, tout en réduisant le risque d'erreurs humaines.Les filtres et les recherches ont automatisé de nombreuses étapes fastidieuses, offrant ainsi un gain de temps et d'efficacité dans le traitement des données. En outre, la vérification approfondie des communes a été réalisée afin de valider leur présence dans la base de référence **HDX**. Les communes mal écrites (exceptés ceux représentées en nombre)ont été corrigées grace à la 2e base. Pour celles manquantes dans le HDX mais présentes dans le EHCVM 2021, une exploraton de cette dernière avec la variable "chef lieu" a permis de confirmer ou modifier le nom de la commune.

Ci-dessous, les sept principales étapes réalisées dans le traitement des données à l'aide de **R** pour obtenir notre base finale.

1. Installations nécessaires

Dans ce projet, nous utilisons plusieurs packages essentiels pour manipuler et analyser les données. Afin de garantir que chaque package est installé uniquement si nécessaire, nous avons défini une fonction *install_if_missing*. Cette fonction vérifie si le package est déjà installé, et dans le cas contraire, l'installe automatiquement.

Voici les packages installés et chargés pour ce projet : *readr*, *readxl*, et *dplyr* (Data Plyr ou dplyr fait référence à un package plus ancien *plyr* pour la manipulation de données). Le code ci-dessous s'assure que ces packages sont disponibles avant d'être utilisés dans le traitement des données :

```
packages <- c("readr", "readxl", "dplyr")
install_if_missing <- function(pkg) {
  if (!requireNamespace(pkg, quietly = TRUE)) install.packages(pkg)
}
invisible(lapply(packages, install_if_missing))

# Charger les packages
library(readr)
```

```
## Warning: le package 'readr' a été compilé avec la version R 4.4.2
```

```
library(readxl)
library(dplyr)
```

```
##
## Attachement du package : 'dplyr'
```

```
## Les objets suivants sont masqués depuis 'package:stats':
##
##     filter, lag

## Les objets suivants sont masqués depuis 'package:base':
##
##     intersect, setdiff, setequal, union
```

2. Chargement des fichiers

Les **chemins relatifs** ont été utilisés pour spécifier les fichiers. L'avantage d'utiliser un chemin relatif plutôt qu'absolu réside dans la portabilité du code, facilitant ainsi l'exécution du script sur différentes machines sans avoir à modifier les chemins d'accès. Cela garantit également que le code fonctionne indépendamment de l'emplacement exact du projet sur l'ordinateur.

```
EHCVM_data <- read_csv("../Data/ehcvm_individu_mli2021.csv",
                       show_col_types = FALSE)
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
# Certaines col ne sont pas bien reconnus
#Certaines colonnes ne sont pas bien reconnues (Par exemple,
#une colonne censée être numérique peut contenir des caractères
#ou bien une colonne attendue comme texte (chr) est vide, ce qui
#cause l'erreur). Comme cela n'affecte pas l'exécution, on masque
#l'avis.

#View(EHCVM_data)

vecteur_data <- read_excel("../Data/Groupe3_tp4_vecteur.xlsx")

#View(vecteur_data)

HDX_data <- read_excel("../Data/mli_adminboundaries_tabulardata.xlsx")
#View(HDX_data)
```

3. Duplication de la var commune dans la base EHCVM

Pour effectuer des comparaisons avvec la base finale à générer, nous avons dupliqué la colonne des communes.

```
# Dupliquer la colonne commune dans EHCVM_datapour conserver
#les communes de EHCVM pour des comparaisons ultérieures

EHCVM_data <- EHCVM_data %>%
  mutate(Commune_EHCVM_dupli = `commune`)

#View(EHCVM_data)
```

4. Première fusion

La première fusion est réalisée entre les données des individus issues de l'EHCVM et le fichier vecteur contenant les informations géographiques des communes. Pour ce faire, nous utilisons la fonction left_join du package dplyr, qui permet d'associer les deux jeux de données sur la base de la variable *Commune_EHCVM_dupli* (présente dans EHCVM_data) et *Admin3_EHCVM* (présente dans vecteur_data). Cette fusion est effectuée dans une relation "many-to-many", ce qui signifie qu'une commune peut apparaître plusieurs fois dans chaque fichier, en fonction des individus qui y sont associés.

```
# Merge 1 : Individus + Fichier Vecteur
merge1 <- EHCVM_data %>%
  left_join(vecteur_data, by = c("Commune_EHCVM_dupli" = "Admin3_EHCVM"),
            relationship = "many-to-many")

#View (merge1)
```

5. Seconde fusion

La deuxième fusion consiste à combiner les résultats de la première fusion avec la base HDX, qui contient des informations administratives supplémentaires. Comme dans la première fusion, nous utilisons à nouveau la fonction left_join de dplyr pour lier les données sur la base de la variable *Admin3_HDX* (dans le jeu de données merge1) et *admin3Name_fr* (dans la base HDX).

Cette fusion permet d'intégrer les informations administratives de la base HDX à notre jeu de données final, offrant ainsi une vue plus complète des communes du Mali.

```
# Merge 2 : Résultat précédents Base HDX
final_data <- merge1 %>%
  left_join(HDX_data, by = c("Admin3_HDX" = "admin3Name_fr"),
            relationship = "many-to-many")

# Vérification du résultat
#View(final_data)
```

6.Vérification

Après avoir effectué les fusions, il est important de vérifier la qualité de la correspondance entre les colonnes *Commune_EHCVM_dupli* et *Admin3_HDX*. Comme la concordance entre les variables commune a été assurée avec le fichier vecteur, nous filtons les lignes où les valeurs des deux colonnes *Commune_EHCVM_dupli* et *Admin3_HDX* ne sont pas manquantes en meme temps.

Le taux de matching donne donc le pourcentage de lignes ayant une correspondance valide.

```
# Vérifier les valeurs non manquantes dans les deux colonnes
matching_rows <- final_data %>%
  filter(!is.na(Commune_EHCVM_dupli) & !is.na(Admin3_HDX)) %>%
  nrow()  # Compter les lignes valides

# Taux de matching en pourcentage
matching_rate <- (matching_rows / nrow(final_data)) * 100

# Afficher
cat("Taux de matching entre EHCVM_Admin3 et HDX_Admin3 :",
    round(matching_rate, 2), "%\n")
```

```
## Taux de matching entre EHCVM_Admin3 et HDX_Admin3 : 79.21 %
```

7. Sauvegarde

Une fois les données fusionnées et vérifiées, il est essentiel de sauvegarder le fichier final pour pouvoir l'utiliser dans d'autres étapes d'analyse. Le code ci-dessous permet de sauvegarder les données fusionnées dans un fichier CSV, que nous avons nommé TP4_G3_FinalMergedData.csv.

```
# Sauvegarde en CSV si besoin
write_csv(final_data, "TP5_G3_FinalMergedData.csv")
```

Conclusion

Ce projet a permis de mettre en évidence l'importance de la préparation des données dans tout processus d'analyse. Grâce à l'utilisation d'**Excel**, nous avons pu gérer précisément les données, en particulier pour résoudre les problèmes de correspondance des noms de communes entre les deux sources de données. Ce travail a été également l'occasion d'appliquer le cours sur le **merging** avec **R**, ce qui nous a permis de mieux comprendre les techniques de fusion de données dans un cadre pratique.

En fin de compte, le fichier enrichi ainsi créé constitue une base solide pour l'analyse géographique des communes du Mali, en offrant des informations fiables et structurées pour enrichir le champs des analyse de l'EHCVM en fucionnant sa base avec les données de HDX.