Partie1_Apurement

KABASSINA Gnimdou Ange & Ibrahima GNING

2022-06-27

Contents

Introduction

Dans les pays en développement, les enquêtes auprès des ménages sont la source privilégiée pour produire des indicateurs de suivi des conditions de vie des populations. c'est dans ce contexte que dans l'espace UEMOA (Union Economique et Monétaire Ouest Africaine) pour le suivi et l'évaluation de la pauvreté et des conditions de vie des ménages dans chacun des pays membres, il a été mis en place l'Enqueête Harmonisée sur les Conditions de vie des Ménages. Et c'est l'analyse des données receuillies dans le cadre de cette enquête au Sénégal qui fait l'objet de cette étude. La présente partie du travail consiste à préparer une table de données propre pour l'analyse : l'appurement.

I. Choix et constitution de la base d'étude

Les données recueillies et mises à notre disposition sont réparties en 20 sections portant chacune sur un aspect de la vie des ménages. Le but de notre travail sera d'étudier le volet éducation contenu dans la section 2 ainsi que quelques caratéristiques sociodémographiques des ménages répertoriés dans la section 1. Ainsi notre base d'étude portera sur certaines variables des section 1 et 2.

I.1 Importation des tables de données des sections 1 & 2

Pour constituer notre base nous utiliserons les informations contenues dans les sections 1 et 2. La prémière étape est donc d'importer les données qui y correspondent.

```
## Appel de la bibliothèque nécessaire pour ouvrir des ficher du format dta
library(haven)

## Définissons le chemin d'accès aux tables de données
path <- "C:\\Users\\HP ProBook\\Documents\\Mes Cours ISEP2\\R\\Projet\\Base de données"

## importation des bases des sections 1 & 2
section_1 <- read_dta(paste(path,"s01_me_SEN_2021.dta",sep="\\")) # tables de la section 1
section_2 <-  read_dta(paste(path,"s02_me_SEN_2021.dta",sep="\\"))# tables de la section 2
```

I.2 Extraction des variables d'intérêt

I.2.1 Caractéristiques Sociodémographiques

Ici nous nous intéresseront à l'étude de:

- L'âge de mariage (pour voir entre autres si les mariages précoses perssistent)
- La nationalité des différents enquêtés (pour voir le dégré d'ouverture du pays c'est à dire s'il y a assez de diversité)
- Les causes des migrations et les localités d'origines des migrants

Pour atteindre ces objectifs on s'interressera à certaines variables à savoir :

- les clés d'identification des ménages : nous les utiliserons pour fusioner les deux sous bases que nous aurons créés-
- le sexe : pour des analyses stratifiées
- Anné de naissance : pour calculer l'âge des enquêtés (pour mieux emputer certaines valeurs manquantes et faire des analyses statifiée)
- Localité d'origine
- Causes de l'installation dans la localité de résidence
- Nationalité

```
# choisissons nos variables d'intérêt, tout en les renommant
## section 1: caractéristiques
carac_men <- section_1[1:6] # variables d'identification tel que la clé du ménage
names(carac_men)[6] <- "member_id"
carac_men[c("Sexe","Situation_Matrimoniale","Age_au_premier_mariage","Nationalite","Ne_dans_sa_localite_
            "Localite_d_origine_des_migrants","Principale_cause_de_migration"
            ,"Annee_de_naissance")] <- section_1[c("s01q01","s01q07","s01q10","s01q15","s01q17", "s01q1
```

I.2.2 Education

Dans le volet éducation, il sera question d'étudier le niveau d'alphabétisation ; les problèmes généralement rencontrés par les élèves et étudiants sénégalais ; les raisons d'abandons et aussi de non scolarisation (on pourra voir par exemple si les gens continuent à penser que les filles n'ont pas le droit de faire des études formelles); l'âge d'entré à l'école... On pourra aussi voir l'impact de la COVID-19 en voyant si les écoles ont fermé et si pendant la ferméture les élèves ont pu rester en contact avec l'administration de l'école ou les enseignants. Ainsi quelques unes des variables qui nous interressent dans la section deux sont:

- Si l'enquêté peut lire, écire, comprendre, un texte en français, en langue locale ou dans une autre langues;

_ Si l'enquêté fait des études formelles;

- Age d'entré à l'école
- problèmes rencontrés à l'école
- Moyens utilisés pour rester en contact avec l'école pendant la fermeture

```
# choisissons nos variables d'intérêt, tout en les renommant
## section 2: Education
edu <- section_2[1:6] # variables d'identification tel que la clé du ménage
names(edu)[6] <- "member_id"
edu[c("Peut_lire_français", "Peut_lire_la_langue_locale", "Peut_lire_une_autre_langue", "Peut_ecrire_fra
edu[c("Peut_comprendre_un_texte_en_français", "Peut_comprendre_un_texte_en_langue_locale", "Peut_compren
edu[c("Fait_une_ecole_formelles","Raisons_de_non_frequentation_d_ecole_formelle","Fait_une_ecole_non_fo
edu[c("Fermeture_de_l_ecole","Reste_en_contact_avec_l_ecole","Resultat_2019_2020","Raisons_d_abandon_de
edu[c("Diplome_le_plus_eleve","Reste_en_contact_avec_l_ecole_par_SMS","Reste_en_contact_avec_l_ecole_pa
edu[c("Insuffisance_de_livres_et_ou_de_fornitures","Insuffisance_de_tables_bancs_et_d_equipements","Abs
```

I.3 Fusion des données

Maintenant que nous avons pu extraires les informations qui nous interressent, nous allons toutes les regrouper en un seul endroit. En d'autre termes on devra à présent fusionner les deux sous bases constituées grâce à la fonction merge.

```
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
```

```r
# fusion des deux sous tables pour avoir notre table de données d'étude
## création de la clé de fusion
carac_men <- mutate(carac_men, key = paste(interview__key,interview__id,id_menage,member_id))
edu <- mutate(edu, key = paste(interview__key,interview__id,id_menage,member_id))
my_data <- merge(carac_men, edu, by = "key")

dim(my_data)
```

```
## [1] 22017    54
```

Notre base d'étude ainsi constituée comprends 22017 lignes et 58 colones. Nous allons à présent y retirer les variables qui se répètent et celle qui ne seront pas trop essentielles à notre analyse. Ensuite nous ajouterons une variable Age dans laquelle nous calculerons l'âge de l'enquêté grâce à son année de naissance. Aussi, les observation en double (doublons) ne nous interresse t-elles pas car elles biaiserons nos analyses ; il faut donc s'en débarasser.

```r
# Supprimons les doublons et les variables en double qui ne nous interressent pas
# c'est à dire les identifiants des ménages
my_data1 <- distinct(my_data) # suppression des doublons
dim(my_data1)
```

```
## [1] 21768    54
```

```r
my_data2 <- my_data1[-c(2:7,16:21)] # suppression des colones ininteressantes

# calculons l'âge des individus
my_data2 <- mutate(my_data2, Age = 2022 - Annee_de_naissance)
```

Les données recueillies sur l'éducation ne concernent que les individus de plus de 3 ans; de même la majeur partie des variables qui nous intérressent et des analyses que nous souhaitons faire, n'ont pas vraiment d'intérêt pour des personnes aussi jeunes (moins de 3 ans). Du coup nous éliminerons cette tranche de la popularion dans nos données.

```r
# excluons les individus de moins de 3 ans car ils ne nous intérressent pas
my_data3 <- subset(my_data2, Age > 3)
dim(my_data3)
```

```
## [1] 13817    43
```

```r
# sauvegardons la base ainsi ontenue dans un ficher csv.
write.csv(my_data3, paste(path,"Base_projet2.csv", sep="\\"))
```

Notre table de données finale comporte donc 13817 lignes et 43 colones ou variables.

II. Recodage ou labélisation des valeurs des variables

Lorsque les données sont importées depuis SAS, SPSS ou Stata, cela permet notamment de conserver le codage original du fichier importé. Mais il faut noter que ces étiquettes de valeur n'indique pas pour autant de manière systématique le type de variable (catégorielle ou continue). Les vecteurs labellisés n'ont donc pas

vocation à être utilisés pour l'analyse, notamment le calcul de modèles statistiques. Ils doivent être convertis en facteurs (pour les variables catégorielles) ou en vecteurs numériques (pour les variables continues).Pour cela nous utiliserons la fonction to_factor de la bibliothèque "labelled"

```
# Recodons les valeurs des variables
# Métode 1 (fastidieuse)
## sexe: 1 = masculin & 2 = féminin
library(dplyr)
my_data4 <- mutate(my_data3, Sexe = if_else(Sexe == 1, "Homme", "Femme", missing = NULL))

# Méthode 2 : transformer en type factor moins fastidieux
# Recodons les valeurs des variables grâce à la fonction to_factor de la librairie "labelled"
library(labelled)
str(my_data3) # pour détecter les variables catégorielles à recoder
```

```
## 'data.frame':    13817 obs. of  43 variables:
##  $ key                                              : chr  "00-01-63-80 1df091103eb748ac859408a
##  $ Sexe                                             : dbl+lbl [1:13817]  1,  2,  1,  1,  1,  1,
##    ..@ label       : chr "1.01. Quel est le sexe de %rostertitle%"
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Masculin" "Féminin"
##  $ Situation_Matrimoniale                           : dbl+lbl [1:13817]  2,  2,  1,  1,  1,  1,
##    ..@ label       : chr "1.07. Quelle est la situation matrimoniale de  %rostertitle%  ?"
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2 3 4 5 6 7
##    .. ..- attr(*, "names")= chr [1:7] "Célibataire" "Marié(e) monogame" "Marié(e) polygame" "Union li
##  $ Age_au_premier_mariage                           : num  25 NA NA NA NA NA 19 17 NA NA ...
##  $ Nationalite                                      : dbl+lbl [1:13817] 13, 13, 13, 13, 13, 13,
##    ..@ label       : chr "1.15. De quelle nationalité est  %rostertitle% ?"
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2 3 4 5 6 7 8 9 10 ...
##    .. ..- attr(*, "names")= chr [1:17] "Bénin" "Burkina Faso" "Cape-vert" "Cote d'ivoire" ...
##  $ Ne_dans_sa_localite_de_residence                 : dbl+lbl [1:13817]  1, NA,  1,  1,  1,  1,
##    ..@ label       : chr "1.17.  %rostertitle%  est-il né au %s01q15%?"
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Localite_d_origine_des_migrants                  : dbl+lbl [1:13817] NA,  1, NA, NA, NA, NA,
##    ..@ label       : chr "1.19. Quelle est la dernière localité où  %rostertitle%  a vécu avant de v
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 4 5 6 7 8 9 10 11 12 ...
##    .. ..- attr(*, "names")= chr [1:20] "Capitale" "Bénin" "Burkina Faso" "Cape-vert" ...
##  $ Principale_cause_de_migration                    : dbl+lbl [1:13817] NA,  9, NA, NA, NA, NA,
##    ..@ label       : chr "1.21. Quelle était la raison principale pour laquelle  %rostertitle%  est
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2 3 4 5 6 7 8 9 10 ...
##    .. ..- attr(*, "names")= chr [1:20] "Envoyé par sa famille pour travailler" "Est venu avec son pa
##  $ Annee_de_naissance                               : num  1983 1997 2001 1991 1991 ...
##  $ Peut_lire_français                               : dbl+lbl [1:13817]  1,  2,  2, NA,  1,  2,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Peut_lire_la_langue_locale                       : dbl+lbl [1:13817]  2,  2,  2, NA,  2,  2,
```

```
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Peut_lire_une_autre_langue                          : dbl+lbl [1:13817]  2, NA,  2, NA,  2,  2,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Peut_ecrire_français                                : num  1 0 0 NA 1 0 0 0 0 0 ...
##  $ Peut_ecrire_la_langue_locale                        : num  NA 0 0 NA 0 0 0 0 0 0 ...
##  $ Peut_ecrire_une_autre_langue                        : num  0 0 0 NA 0 0 0 0 0 NA ...
##  $ Peut_comprendre_un_texte_en_français                : num  1 0 0 NA 1 0 0 0 0 0 ...
##  $ Peut_comprendre_un_texte_en_langue_locale           : num  0 0 0 NA 0 0 0 NA NA 0 ...
##  $ Peut_comprendre_un_texte_en_autre_langue            : num  0 0 0 NA 0 0 0 0 NA 0 ...
##  $ Fait_une_ecole_formelles                            : dbl+lbl [1:13817]  1,  2, NA, NA,  1, NA,
##    ..@ label       : chr "2.03. %rostertitle% a-t-il fait ou fait-il des études actuellement dans une
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Raisons_de_non_frequentation_d_ecole_formelle       : dbl+lbl [1:13817] NA, 14, 14, NA, NA, 14,
##    ..@ label       : chr "2.04. Pour quelle raison principale %rostertitle% n'a-t-il pas fait des étu
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2 3 4 5 6 7 8 9 10 ...
##    .. ..- attr(*, "names")= chr [1:15] "Trop jeune" "Pas d'école, école trop éloignée" "Refus de la f
##  $ Fait_une_ecole_non_formelle                         : dbl+lbl [1:13817] NA,  1,  1, NA, NA,  1,
##    ..@ label       : chr "2.05. Est ce que %rostertitle% a suivi  une école non formelle ou une forma
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Age_d_entre_a_l_ecole                               : num  6 NA NA NA 7 NA NA NA NA NA ...
##  $ Gerant_de_l_ecole                                   : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ label       : chr "2.09. Qui gère l'école fréquentée par %rostertitle% au cours de l'année %an
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2 3 4 5 6
##    .. ..- attr(*, "names")= chr [1:6] "Gouvernement" "Privé religieux" "Privé non religieux" "Privé n
##  $ Fermeture_de_l_ecole                                : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ label       : chr "2.09a. L'école de  %rostertitle%  a-t-elle été fermée momentanément à cause
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Reste_en_contact_avec_l_ecole                       : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ label       : chr "2.09b.  %rostertitle%  était-il/elle en contact avec les enseignants ou l'a
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Resultat_2019_2020                                  : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ label       : chr "2.10. Quel résultat %rostertitle% a-t-il obtenu au cours de l'année %annee$
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2 3 4 5
##    .. ..- attr(*, "names")= chr [1:5] "Diplômé, études achevées" "Passe en classe supérieure" "Echec
##  $ Raisons_d_abandon_de_l_ecole                        : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ label       : chr "2.11. Pour quelle raison %rostertitle% a-t-il abandonné l'école en cours d
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2 3 4 5 6 7 8 9 10 ...
##    .. ..- attr(*, "names")= chr [1:16] "A obtenu un emploi" "S'est marié" "C'est une fille" "Grossess
```

```
##  $ Diplome_le_plus_eleve                                  : dbl+lbl [1:13817]  0, NA, NA, NA,  1, NA,
##    ..@ label       : chr "2.33. Quel est le diplôme le plus elevé obtenu par %rostertitle%?"
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  0 1 2 3 4 5 6 7 8 9 ...
##    .. ..- attr(*, "names")= chr [1:11] "Aucun" "CEPE" "BEPC" "CAP" ...
##  $ Reste_en_contact_avec_l_ecole_par_SMS                  : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Reste_en_contact_avec_l_ecole_par_applications_mobiles: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Reste_en_contact_avec_l_ecole_par_Email                : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Reste_en_contact_avec_l_ecole_par_courier              : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Reste_en_contact_avec_l_ecole_Telephone                : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Insuffisance_de_livres_et_ou_de_fornitures             : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Insuffisance_de_tables_bancs_et_d_equipements          : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Absenteisme_des_enseignants_greve                      : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Enseignement_pas_satisfaisant                          : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Effectifs_plethoriques                                 : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Insuffisance_d_enseignants                             : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Manque_de_toilettes                                    : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Frequence_des_cotisations                              : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Salle_de_classe_en_mauvais_etat                        : dbl+lbl [1:13817] NA, NA, NA, NA, NA, NA,
##    ..@ format.stata: chr "%10.0g"
##    ..@ labels      : Named num  1 2
##    .. ..- attr(*, "names")= chr [1:2] "Oui" "Non"
##  $ Age                                                    : num  39 25 21 31 31 26 47 45 6 7 ...
```

```
# variables à recoder
variables = c("Sexe", "Situation_Matrimoniale", "Nationalite","Ne_dans_sa_localite_de_residence"
              ,"Localite_d_origine_des_migrants","Principale_cause_de_migration","Fait_une_ecole_formell
              ,"Raisons_de_non_frequentation_d_ecole_formelle","Fait_une_ecole_non_formelle","Gerant_de_
              'Fermeture_de_l_ecole',"Reste_en_contact_avec_l_ecole","Resultat_2019_2020","Raisons_d_aba
              ,"Diplome_le_plus_eleve","Insuffisance_de_livres_et_ou_de_fornitures","Insuffisance_de_tab
```

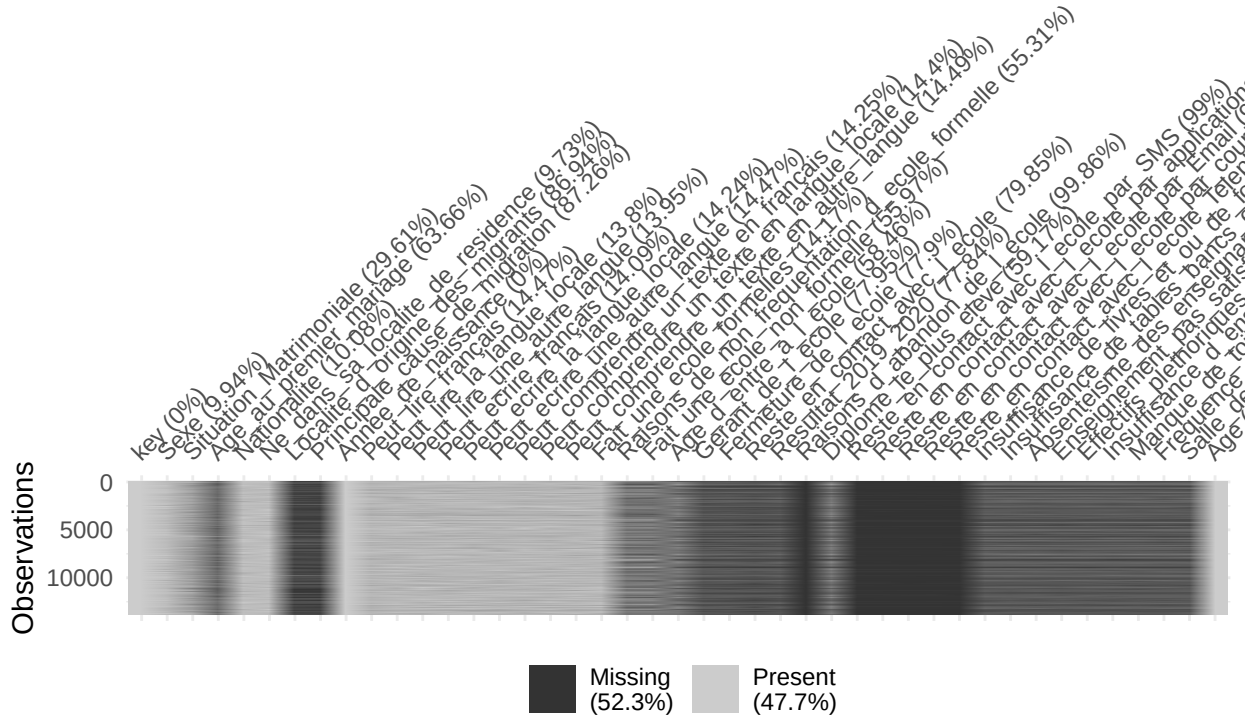
```
                ,'Absenteisme_des_enseignants_greve',"Enseignement_pas_satisfaisant","Effectifs_plethoriqu
                "Manque_de_toilettes",'Frequence_des_cotisations',"Salle_de_classe_en_mauvais_etat",
                "Peut_lire_français","Peut_lire_la_langue_locale","Peut_lire_une_autre_langue")

my_data4[variables] <- to_factor(my_data3[variables]) # recodage
```

III. Gestion des valeurs manquantes

Le but de cette partie est d'imputer les variables manquantes grâce à certaines méthodes que nous verrons plus tard. Il faut cependant noter que certaines valeurs manquantes sont normales car les questions auquelles elles font référence ne consernent pas le répondant. pour ce cas de figure, on conservera les valeurs manquantes ou les changerons par 0 = "pas concerné" pour être plus précis. Ces contrôles seront faits dans la partie suivante qui a pour but de détecter et gérer les valeurs aberrantes.

Avant d'appliquer nos méthodes d'imputation, nous allons visualiser les données manquantes afin de voir leur pourcentage dans nos données.

```
## visualisation générale des valeurs manquantes

library(visdat)
vis_miss(my_data4)
```



Ce graphique nous montre qu'il n'y a presque pas d'informations sur les moyens utilisés pour rester en contact avec l'école pendant sa fermuture suite à la covid-19 (pourcentage de NA supérieur à 90%). Essayer

d'emputer ces NA reviendrait pratiquement à créer nos propres observations ce qui dangereux. Ainsi nous exclurons tout simplement le volet 'Impact de la covid sur le fonctionement des écoles' dans notre analyse.

```
covid <- c("Fermeture_de_l_ecole","Reste_en_contact_avec_l_ecole","Reste_en_contact_avec_l_ecole_par_SMS
my_data4 <- select(my_data4,- covid)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(covid)` instead of `covid` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

- Facultées de maîtrise d'une langue :

Une des analyses qui sera portée sur ces données sera de savoir globalement est-ce que les individus maîtrisent le français par exemple. Pour cela les valeurs manquantes de ces variables (catégorielles) seront remplacer par le mode parce que le mode représente la modalité avec la plus grande probabilité d'observation ou d'appiration. En d'autre terme si on pose une question à un individu, il y a une plus grande chance que sa réponse soit le mode ce qui nous permettra d'évaluer la situation la plus probable en terme de maîtrise des langues.

```
## Variables de maîtrise de la langue
langue <- c("Peut_lire_français","Peut_lire_la_langue_locale","Peut_lire_une_autre_langue",
            "Peut_ecrire_français", "Peut_ecrire_la_langue_locale", "Peut_ecrire_une_autre_langue",
            "Peut_comprendre_un_texte_en_français","Peut_comprendre_un_texte_en_langue_locale",
            "Peut_comprendre_un_texte_en_autre_langue")

### cherchons le mode et detection des na et remplacement par le mode
for (x in langue){
  y <- sort(table(my_data4[x]))# on range les modalité en fonction de leurs fréquence dans l'ordre croi
  print(y) # pour voir les différentes modalités et le mode
  y <- data.frame(y)
  na <- which(is.na(my_data4[x])) # detection des valeurs manquantes
  my_data4[na,x] <- y[2,1] # remplacement des na
}
```

```
## Peut_lire_français
##  Oui  Non
## 5678 6139
## Peut_lire_la_langue_locale
##   Oui   Non
##   574 11336
## Peut_lire_une_autre_langue
##  Oui  Non
## 2258 9632
## Peut_ecrire_français
##    1    0
## 5460 6410
## Peut_ecrire_la_langue_locale
##     1     0
##   548 11302
## Peut_ecrire_une_autre_langue
##    1    0
## 2195 9623
```

```
## Peut_comprendre_un_texte_en_français
##    1    0
## 5223 6625
## Peut_comprendre_un_texte_en_langue_locale
##     1     0
##   955 10872
## Peut_comprendre_un_texte_en_autre_langue
##    1    0
## 2026 9789
```

- Fréquentation d'une école ou non et raisons de non fréquentation

Pour les valeurs manquantes à ce niveau, elle seront remplacées par le mode comme précédement.

```
formation <- c("Fait_une_ecole_formelles","Raisons_de_non_frequentation_d_ecole_formelle","Fait_une_ecol
for (x in formation){
  y <- sort(table(my_data4[x]))# on range les modalité en fonction de leurs
  # fréquence dans l'ordre croissant. Ainsi le mode est le premier élément de y
  y <- data.frame(y)
  na <- which(is.na(my_data4[x])) # detection des valeurs manquantes
  my_data4[na,x] <- y[2,1] # remplacement des na
}
```

- Age d'entrée à l'école

La gestion des valeur manquantes pour cette variable consitera en leur remplacement par la médiane car cette dernière est moins sensibles aux valeurs abérantes.

```
## gestion des NA pour l'âge d'entrée à l'école

x <- which(is.na(my_data4["Age_d_entre_a_l_ecole"])) # detection des Na
my_data4[x,"Age_d_entre_a_l_ecole"] <- median(my_data4[which(is.na(my_data4["Age_d_entre_a_l_ecole"])==
```

- Principales difficultées rencontrées à l'école :

A ce niveau, nous utiliserons une aproche puissante celle des K plus proches voisins. Le principe de cette méthode est simple et assez intuitif. Il se base sur l'hypothèse selon laquelle les individus appartenant à un même groupe ont tendance à avoir des comportements similaires c'est à dire à présenter les mêmes. En d'autre termes pour remplacer la valeur manquante d'un individus, on constitue un groupe des K individus les plus proche de ce individu en termes d'attributs ou caractéristiques et on remplace la valeur manquant par la moyenne des observations pour une variable numérique et par le mode (vote majoritaire) s'il s'agit de variables catégorielles. Cette approche est implémenté dans R grâce à la fonction "KNN" de la bibliothèque "VIM"

```
## gestion des variables relatives aux difficultés rencontrés à l'école

library(VIM)
```

```
## Le chargement a nécessité le package : colorspace
```

```
## Le chargement a nécessité le package : grid
```

```
## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attachement du package : 'VIM'

## L'objet suivant est masqué depuis 'package:datasets':
##
##     sleep
```

```
dificulties <- c("Insuffisance_de_livres_et_ou_de_fornitures","Insuffisance_de_tables_bancs_et_d_equipe
                 ,"Absenteisme_des_enseignants_greve", "Enseignement_pas_satisfaisant", "Effectifs_plet
                 "Insuffisance_d_enseignants", "Manque_de_toilettes", "Frequence_des_cotisations", "Sal
## cherchons toutes les lignes qui présentes des NA pour ces variables
dificulties_na <- c() # vecteur qui recueillera les index de ces lignes
for (x in dificulties){
  x <- which(is.na(my_data4[x])) # detection des valeur maquantes
  dificulties_na <- c(dificulties_na, x) # ajout des index aux autres précédement déterminés
}
y = data_frame(dificulties_na)
dificulties_na <- dificulties_na[!duplicated(y)] # supprimer les doublons car on ne veux pas que les in
data_s = my_data4[10:36] # variables portant sur l'éducation
my_data4[dificulties_na,10:36] <- kNN(data_s[dificulties_na,]) # imputation
```

- Sexe

Au Sénégal, comme dans plusieurs pays de la sous région, on a plus de femmes que d'hommes. En effet, selon les projections démographiques de l'Agence nationale de la Statistique et de la Démographie (ANSD), en 2017 , la population sénégalaise était estimée à 15 256 356 habitants dont 50,2% de femmes. C'est de ce constat général que part notre méthode d'imputation de valeurs manquantes qui consiste à remplacer les NA pour la variable sexe par "Féminin"

```
## Variable Sexe
x <- which(is.na(my_data4["Sexe"])) # detection des valeur maquantes
my_data4[x,"Sexe"] <- "Féminin"  # imputation
```

- Nationalité

En se référant à ce qui a été fait précédement pour le sexe, nous remplacerons les nationnalités manquantes par celle sénégalaises car, le constat est que, nous nous trouvons sur le territoire sénégalais et donc la majeure partie des résidants ne peuvent être que des sénégalais.

```
## variable Nationalité
x <- which(is.na(my_data4["Nationalite"])) # detection des valeur maquantes
my_data4[x,"Nationalite"] <- "Sénégal"  # imputation
```

- Autres variables de caractéristiques sociodémographiques

Ici nous allons poursuivre dans la même lancée en utilisant une méthode qui va remplacer le données manquantes par une prédiction de qu'elle aurait pu être en fonction de ces autres attributs. le modèle de prédiction que utiliserons sera l'arbre de décision. Cela sera possible grâce à la fonction mice() de la bibliothèque MICE

```r
demographie <- c("Situation_Matrimoniale", "Age_au_premier_mariage",
                 "Ne_dans_sa_localite_de_residence","Localite_d_origine_des_migrants","Principale_cause_

demographie_na <- c() # vecteur qui recueillera les index de ces lignes
for (x in demographie){
  x <- which(is.na(my_data4[x])) # detection des valeur maquantes
  demographie_na <- c(demographie_na, x) # ajout des index aux autres précédement déterminés
}
y = data_frame(demographie_na)
demographie_na <- demographie_na[!duplicated(y)] # supprimer les doublons car on ne veux pas que les in

library(mice)
```

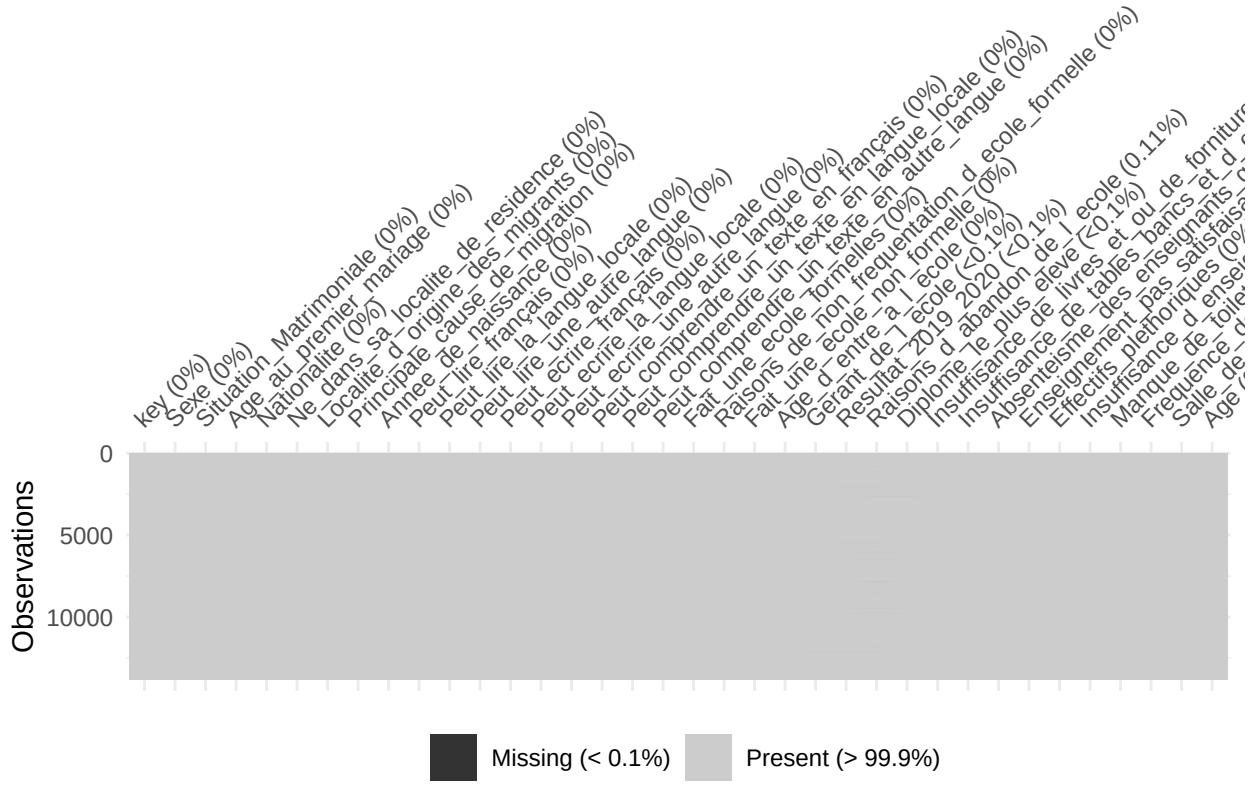
```
##
## Attachement du package : 'mice'

## L'objet suivant est masqué depuis 'package:stats':
##
##     filter

## Les objets suivants sont masqués depuis 'package:base':
##
##     cbind, rbind
```

```r
my_data4[demographie_na,] <-  complete(mice(my_data4[demographie_na,],meth = "cart"),1) # imputation
```

```
##
##  iter imp variable
##   1   1  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   1   2  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   1   3  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   1   4  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   1   5  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   2   1  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   2   2  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   2   3  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   2   4  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   2   5  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   3   1  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   3   2  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   3   3  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   3   4  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   3   5  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   4   1  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   4   2  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   4   3  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   4   4  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   4   5  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   5   1  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   5   2  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   5   3  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   5   4  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
##   5   5  Situation_Matrimoniale  Age_au_premier_mariage  Ne_dans_sa_localite_de_residence  Localite_
```

Vérifions maintenant si toutes nos valeurs manquantes ont bien été remplacées.

```
vis_miss(my_data4)
```




```
basef <- my_data4
```

Super ! toutes nos valeurs manquantes ont disparues ! Mais cela garantit-il la qualité de nos données ? Toutes les observations sont-elles cohérantes ? C'est ce que nous verrons dans la partie suivante.

IV. Gestion des valeurs aberrantes

Comme brièvement abordé par un exemple dans la partie précédente, le but de la présente partie est de gérer les valeurs qui n'ont pas de sens.On s'interressera aussi aux valeurs aberrantes qui sont des valeurs ou observations qui sont distantes des autres observations effectuées sur le même phénomène.

- Age du premier mariage

Comme son nom l'indique, cette variable est sensée nous donner l'âge qu'avait l'interressé lors de sont premier mariage. Il est donc clair que la réponse d'un célibataire à cette question ne peut être valide (doit être NA). Par aillers, lors de la collecte de données, l'agent enquêteur a du remplir 9999 pour certains individus pour marquer que la réponse était inconue. Il faudra également en tenir compte.

```r
## Age au premier mariage
## les célibataires ne sont pas consernés
y <- which(basef["Situation_Matrimoniale"] == "Célibataire" | basef["Age_au_premier_mariage"] == 9999)
basef[y,"Age_au_premier_mariage"] <- NA
```

variables portant sur la migration

les données sur la migration étant assez semblibes, nous conserverons les données d'origines pour pouvoir faire des analyses plus pertinantes

```r
## conserver les données d'origines pour faire des analyses plus pertinantes
migr <- c("Localite_d_origine_des_migrants","Principale_cause_de_migration")
basef[migr] <- to_factor(my_data3[migr])
```

- Facultées de maîtrise d'une langue :

ici nous alons juste vérifier les modalités ; a t-on bien 2 modalités ou plus ou moins ? ces dernières sont-elles conforment au questionnaire ?

```r
## Variables de maîtrise de la langue
### Gestion valeurs aberrantes
### jettons un coup d'oeil sur les modalités de nos variables
### pour voir si tout va bien

for(x in langue){
  print(table(basef[x]))
}
```

```
## Peut_lire_français
##  Oui  Non
## 5678 8139
## Peut_lire_la_langue_locale
##   Oui   Non
##   574 13243
## Peut_lire_une_autre_langue
##   Oui   Non
##  2258 11559
## Peut_ecrire_français
##    0    1    2
## 6410 5460 1947
## Peut_ecrire_la_langue_locale
##     0     1     2
## 11302   548  1967
## Peut_ecrire_une_autre_langue
##    0    1    2
## 9623 2195 1999
## Peut_comprendre_un_texte_en_français
##    0    1    2
## 6625 5223 1969
## Peut_comprendre_un_texte_en_langue_locale
##     0     1     2
```

```
## 10872   955  1990
## Peut_comprendre_un_texte_en_autre_langue
##    0    1    2
## 9789 2026 2002
```

On voit qu'il y a 3 modalité au lieu de 2 pour certaines variables. De plus nous avons vu depuis l'imputation que ces dernières n'étaient pas conforment au questionaire. on va donc régler tout ça. la modalité "2" provient de l'imputation et devrait plutôt être égal à 0 le mode (d'après les tableaux de fréquences). il y a donc eu un soucis et 0 doit être remplacer par 2 pour rester conforme au questionnaire.

```
langue2 <- c("Peut_ecrire_français", "Peut_ecrire_la_langue_locale", "Peut_ecrire_une_autre_langue",
             "Peut_comprendre_un_texte_en_français","Peut_comprendre_un_texte_en_langue_locale",
             "Peut_comprendre_un_texte_en_autre_langue")
for(x in langue2){
  y <- which(basef[x] == 0)
  basef[y,x] <- 2
}
```

- Fréquentation d'une école ou non et raisons de non fréquentation

Cette partie porte sur les variables qui nous renseigne si l'individu fait une école formelle ou non; sinon pourquoi. alors un premier contrôle porte sur les raisons de non fréquentation d'une école formelle car, pour cette question, un individu qui fait déjà une formation formelle n'est pas concerné. Doù une parie des na provienne du fait que l'individus fait effectivement une formation formelle, ainsi pour de tels enquêtés, la reponse à cette question sera remplacée par NA.

```
## vérification de réponses 2: faire en sorte que la réponse à la question "quels sont les raisons de n

x <- which(basef["Fait_une_ecole_formelles"] == "Oui") # detecter les observations où les individus fon
basef[x,"Raisons_de_non_frequentation_d_ecole_formelle"] <- NA # pour chaque individu (ligne) identifié
```

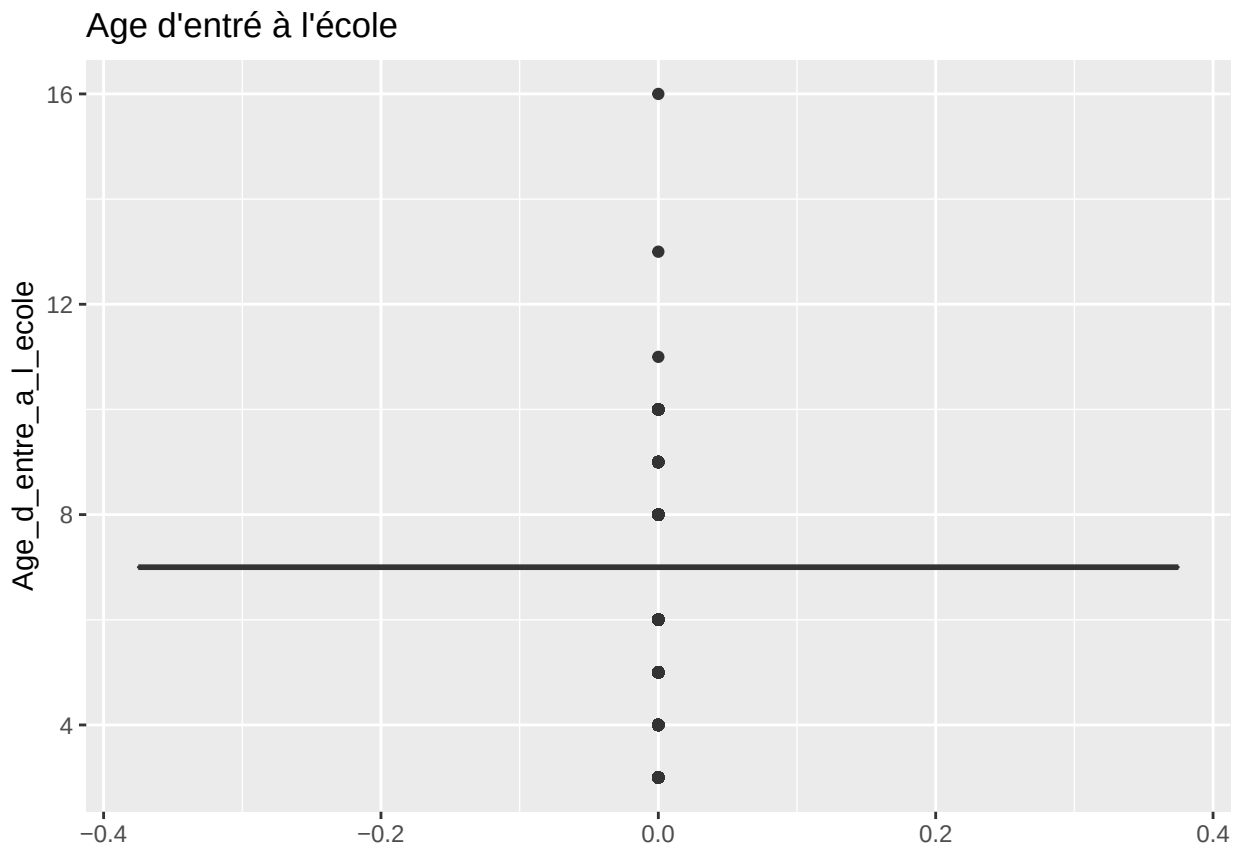
- Age et âge d'entrée à l'école

un premier contrôle à effectuer consiste à s'assurer que tous ce qui n'ont pas fait de formation ne renseignent l'âge d'entré à l'école.

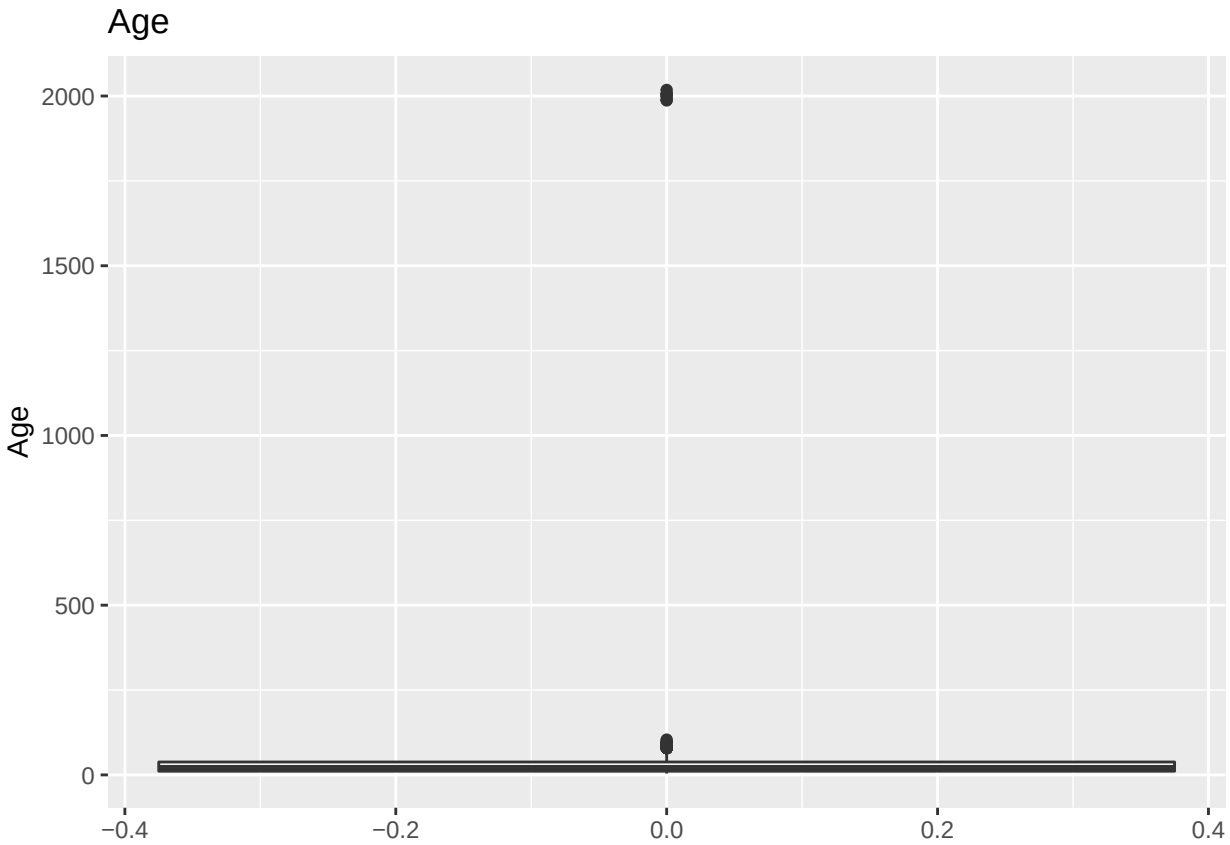
```
## ceux qui n'ont fait aucune formation ne doivent renseigner cette partie
y <- which(basef["Fait_une_ecole_formelles"] == "Non" & basef["Fait_une_ecole_non_formelle"] == "Non")
basef[y,"Age_d_entre_a_l_ecole"] <- NA
```

Pour le second contrôle, consiste à voir s'il y a des outliers. A cet effet, une méthode classiquement employée pour détecter les outliers (ou donnée aberrante), consiste à réaliser un boxplot.

```
library(ggplot2)
ggplot(basef, aes( y=Age_d_entre_a_l_ecole,fill=Age_d_entre_a_l_ecole)) +
  geom_boxplot()+
  ggtitle("Age d'entré à l'école")
```




```
## on fait pareil pour l'age
ggplot(my_data4, aes( y=Age,fill=Age)) +
  geom_boxplot()+
  ggtitle("Age")
```



Sur cette visualisation des données, les outliers sont représentés sous forme de points. Ils correspondent à des observations dont les valeurs sont :

- supérieures à la valeur du 3ème quartile plus 1.5 fois l'intervalle inter-quartile,
- ou inférieures à la valeur du 1er quartile moins 1.5 fois l'intervalle inter-quartile.

Pour récupérer ces valeurs aberrantes, on utilise la fonction boxplot.stats et avec la fonction which que nous avons déjà beaucoup utilisé, nous pourrons recueillir les index de ces valeurs pour les remplacer par la moyenne.

```
### recupérer les valeurs aberrantes
outlier <- boxplot.stats(basef$Age_d_entre_a_l_ecole)$out
### recupérer les positions des outliers
outlier_ind <- which(basef$Age_d_entre_a_l_ecole %in% c(outlier))
### remplacer les outliers par la moyenne
basef[outlier_ind,"Age_d_entre_a_l_ecole"] <- mean(basef$Age_d_entre_a_l_ecole)

### recupérer les valeurs aberrantes
outlier <- boxplot.stats(basef$Age)$out
### recupérer les positions des outliers
outlier_ind <- which(basef$Age %in% c(outlier))
### remplacer les outliers par la médiane
basef[outlier_ind,"Age"] <- median(basef$Age)
```

- Raison d'abandon des classes et de non fréquentation

Il faut s'assurer pour ces rebriques que les reponses "c'est une fille" ou "Grossesse" correspondent bien à des individus de sexe féminin.

```
## Raison d'abandon des classes et de non fréquentation d'école formelle
## cherchons les hommes qui ont pour raison d'abandon des classe ou de non fréquentation "c'est une fil
## remplaçons la raison par NA
### raison de non fréquentation
x1 <- which(basef["Sexe"] == "Masculin" & basef["Raisons_de_non_frequentation_d_ecole_formelle"] == "C'
basef[x1,"Raisons_de_non_frequentation_d_ecole_formelle"] <- NA
### raisons d'abandon; on gèrera aussi les cas des homme ayant pour raison "grossesse"
x <- which(basef["Sexe"] == "Masculin" & basef["Raisons_d_abandon_de_l_ecole"] == "C'est une fille")
x2 <- which(basef["Sexe"] == "Masculin" & basef["Raisons_d_abandon_de_l_ecole"] == "Grossesse")
basef[x,"Raisons_d_abandon_de_l_ecole"] <- NA
basef[x2,"Raisons_d_abandon_de_l_ecole"] <- NA
```

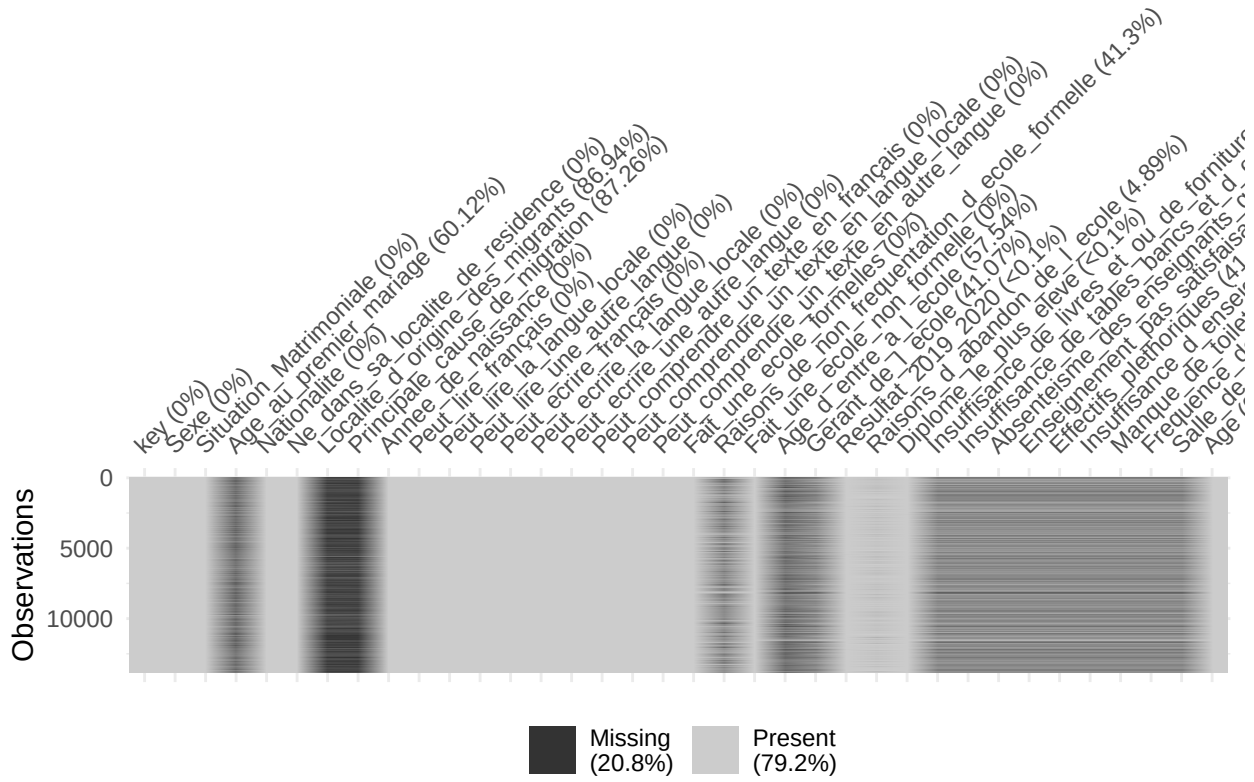
- Gérant de l'école et difficultés rencontrés dans le cadre de la formation

Cette question s'adresse aux individus qui vont actuellement à l'école ainsi un répondant qui ne va ni dans une école formelle ni dans une école informelle ne devrait y répondre. Nous remplaçerons ces réponses par des NA. Pareil pour les difficultés rencontrés dans le cadre de la formation.

```
## Gérant de l'école: la réponse de tous ceux qui ne font aucune école doit être NA pareil pour les dif
var <- c("Gerant_de_l_ecole",dificulties)
for(x in var){
  y <- which(basef["Fait_une_ecole_formelles"] == "Non" & basef["Fait_une_ecole_non_formelle"] == "Non"
  basef[y,x] <- NA
}
```

Puisque nous avons introduit des NA dernièrement, visualusons les une dernière fois puis exportons notre table de données d'étude finale.

```
# visualisation des NA sur nos données finales
base_finale <- basef
vis_miss(base_finale)
```




```
# sauvegardons la base ainsi ontenue dans un ficher dta et un autre csv.
# mais avant il faut changer les noms qui sont trop longs pour pouvoir les exporter en format .dta
name <- c("Peut_comprendre_un_texte_en_français", "Peut_comprendre_un_texte_en_langue_locale",
          "Peut_comprendre_un_texte_en_autre_langue", "Raisons_de_non_frequentation_d_ecole_formelle"
          ,"Insuffisance_de_livres_et_ou_de_fornitures","Insuffisance_de_tables_bancs_et_d_equipements"
          ,"Absenteisme_des_enseignants_greve")
new_names <- c("comprendre_un_texte_en_français", "comprendre_texte_langue_locale",
               "comprendre_texte_en_autre_langue", "Raisons_non_frequentation_EF"
               ,"Insuffisance_livres_fornitures","Insuffisance_tablesb_equipements"
               ,"Absenteisme_enseignants_greve")

for(i in 1:length(name)){
  names(base_finale)[names(base_finale) == name[i]] = new_names[i]
}

write.csv(base_finale, paste(path,"Base_projetf.csv", sep="\\"))

write_dta(base_finale, paste(path,"Base_projetf.dta", sep="\\"))
```

La première partie prend ainsi fin ! Passons à l'analyse de nos données.