Untitled

Fallou BADJI

2023-07-24

Contents


```
knitr::opts_chunk$set(echo = TRUE)
```

AGENCE NATIONALE DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE

ÉCOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE ÉCONOMIQUE

Projet sur le logiciel R

Fallou BADJI
Élève Ingénieur Statisticien Économiste
Chargé du cours : M. Aboubacar HEMA

Le 24 juillet 2023

Introduction

L'objectif de ce projet est que nous appliquions les outils que nous avez étudiés dans le cours du logiciel statistique R, dans le cas d'une étude de cas réelle. Le devoir est à faire seul et à rendre au format .docx ou .pdf. Les codes que nous utiliserons pour répondre aux questions seront à intégrer dans le corps de notre rapport. Nous nous penchons sur R Markdown pour mener à bien notre projet. En ce qui concerne l'organisation du travail à rendre, nous nous inspirerez de la façon dont est organisé le sujet du projet.

Partie 1

Cette enquête vise à identifier et à caractériser des bioénergies durables pour les petites et moyennes entreprises (PME) agroalimentaires d'Afrique de l'Ouest.

Préparation des données

Le fichier Base_Partie1.xlsx contient 250 observations et 33 variables. La première colonne key correspond à l'identifiant de la PME.

Importation et mise en forme

Nous commençons par importer la base "Base_Partie1.xlsx" qui est de type Excel.

```
library(readxl) # pour importer les fichiers de types Excel
# importation de la base dans un objet nommé projet de type dataframe
projet <- as.data.frame(readxl::read_excel("Base_Partie 1.xlsx"))
```

Faisons un tableau qui resume les valeurs manquantes par variable

```
#On recupère le nombre de valeurs manquantes pour chaque variable
t <- sapply(projet,function(x) sum(is.na(x)))
#On les range dans une matrice de taille (11,3) pour former un tableau qui ne prend pas de longueur
Nbre_NA <- matrix(t, ncol= 11)
# On prend les noms des variables
Var <- matrix(names(projet), ncol = 11)
# On affecte chaque valeurs à son nombre de valeurs manquantes
data.frame(Var[1,], Nbre_NA[1,], Var[2,], Nbre_NA[2,], Var[3,], Nbre_NA[3,])
```

```
##          Var.1... Nbre_NA.1...        Var.2... Nbre_NA.2...         Var.3...
## 1             key            0              q1            0               q2
## 2             q23            0             q24            0            q24a_1
## 3          q24a_2            0          q24a_3            0            q24a_4
## 4          q24a_5            0          q24a_6            0            q24a_7
## 5          q24a_9            0         q24a_10            0              q25
## 6             q26            0             q12            0             q14b
## 7             q16            1             q17          131              q19
## 8             q20            0       filiere_1            0        filiere_2
## 9       filiere_3            0       filiere_4            0               q8
## 10            q81            0 gps_menlatitude            0 gps_menlongitude
## 11 submissiondate            0           start            0            today
```

```
##    Nbre_NA.3...
## 1             0
## 2             0
## 3             0
## 4             0
## 5             0
## 6             1
## 7           120
## 8             0
## 9             0
## 10            0
## 11            0
```

Vérifions s'il y a des valeurs manquantes pour la variable key dans la base projet:

```
print("Le nombre de valeurs manquantes pour la variable key est :")
```

```
## [1] "Le nombre de valeurs manquantes pour la variable key est :"
```

```
sum(is.na(projet$key)) #somme des valeurs manquantes sur la variable key
```

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

Création de variables

Rénommons la variable q2 en departement et la variable q23 en sexe. Puis créons la variable sexe_2 qui vaut 1 si sexe égale à Femme et 0 sinon. ENsuite, Créons un data.frame nommé langues qui prend les variables key et les variables concernant les langues.

```
library(dplyr) #pour apporter des modifications sur la base comme changer les noms des variables on sel
# On selectionne les variables à renommer en leur donnant directement et respectivement leurs nouveaux
projet <- projet %>% dplyr::rename(region = q1, departement = q2, sexe = q23)
#On crée une nouvelle variable sexe_2 qu'on ralie à la base
projet <- dplyr::mutate(projet,sexe_2 = ifelse(sexe == "Femme", "1", "0"))
#On merge les dataframes avec la fonction merge.data.frame
langues <- merge.data.frame(projet$key, projet %>% dplyr::select(contains("q24a_"))) #On choisit les va
head(langues)
```

```
##                                         x q24a_1 q24a_2 q24a_3 q24a_4 q24a_5
## 1 uuid:68bff42b-1228-4c66-9bcc-e6d312d9fea6      0      1      0      1      0
## 2 uuid:d70b3c7e-3ca0-4358-bc59-3f7f6baf55e9      0      1      0      1      0
## 3 uuid:0ac18b64-7d85-4bb9-a842-698ac79909af      0      1      0      1      0
## 4 uuid:c52cf5e4-8c28-4e65-998b-3fe2a971a1a3      0      1      0      1      0
## 5 uuid:ac177870-001c-4ada-8747-c22ffe4e4596      0      1      0      1      0
## 6 uuid:578097cf-9af7-46e6-8992-d9079b14c342      0      1      0      1      0
##   q24a_6 q24a_7 q24a_9 q24a_10
## 1      0      0      0       0
## 2      0      0      0       0
## 3      0      0      0       0
## 4      0      0      0       0
## 5      0      0      0       0
## 6      0      0      0       0
```

Créons une variable parle qui est égale au nombre de langue parlée par le dirigeant de la PME. Sélectionnons uniquement les variables key et parle, l'objet de retour sera langues. Et enfin, mergeons les data.frame projet et langues.

```
parle <- rowSums(projet %>% dplyr::select(contains("q24a_")))
langues <- data.frame(key = projet$key,parle)
projet <- dplyr::left_join(projet,langues, by = "key")
head(projet)
```

```
##                                     key     region departement  sexe q24
## 1 uuid:68bff42b-1228-4c66-9bcc-e6d312d9fea6   Diourbel      Bambey Femme  65
## 2 uuid:d70b3c7e-3ca0-4358-bc59-3f7f6baf55e9      Thiès       Mbour Femme  52
## 3 uuid:0ac18b64-7d85-4bb9-a842-698ac79909af      Thiès       Mbour Femme  65
## 4 uuid:c52cf5e4-8c28-4e65-998b-3fe2a971a1a3      Thiès       Mbour Femme  38
## 5 uuid:ac177870-001c-4ada-8747-c22ffe4e4596 Ziguinchor     Bignona Homme  40
## 6 uuid:578097cf-9af7-46e6-8992-d9079b14c342 Ziguinchor    Oussouye Femme  43
##   q24a_1 q24a_2 q24a_3 q24a_4 q24a_5 q24a_6 q24a_7 q24a_9 q24a_10
## 1      0      1      0      1      0      0      0      0       0
## 2      1      1      0      0      1      0      0      0       0
## 3      1      1      0      0      0      0      0      0       0
## 4      1      1      0      0      1      0      0      0       0
## 5      1      1      1      0      0      1      0      0       0
## 6      1      1      1      0      0      0      0      0       0
##                 q25 q26 q12 q14b q16        q17          q19 q20 filiere_1
## 1      Aucun niveau  40 GIE  Non Non       <NA> Mauvais état Oui         1
## 2      Aucun niveau   3 GIE  Non Non       <NA>   Etat moyen Oui         1
## 3   Niveau primaire   5 GIE  Non Non       <NA>   Etat moyen Oui         1
## 4   Niveau primaire   2 GIE  Non Oui Etat moyen         <NA> Oui         0
## 5 Niveau secondaire  20 GIE  Oui Oui   Bon état         <NA> Oui         0
## 6 Niveau secondaire  15 GIE  Non Oui   Bon état         <NA> Oui         0
##   filiere_2 filiere_3 filiere_4                            q8          q81
## 1         0         0         0                      Aucun Propriétaire
## 2         0         1         0 Tansformation d'autres céréales Propriétaire
## 3         0         1         0 Tansformation d'autres céréales Propriétaire
## 4         0         1         1 Tansformation d'autres céréales Propriétaire
## 5         0         0         1               Autre a preciser Propriétaire
## 6         0         0         1     Transformation de la mangue Propriétaire
##   gps_menlatitude gps_menlongitude      submissiondate               start
## 1        14.62691        -16.46786 2021-06-14 20:04:38 2021-06-14 15:38:19
## 2        14.39973        -16.95614 2021-06-07 21:58:11 2021-06-03 19:55:41
## 3        14.39813        -16.95576 2021-06-03 19:33:53 2021-06-03 16:52:49
## 4        14.40838        -16.96077 2021-06-12 16:57:28 2021-06-12 16:20:54
## 5        13.04763        -16.61266 2021-05-17 13:56:10 2021-05-10 13:18:52
## 6        12.48649        -16.54560 2021-06-14 15:27:34 2021-06-11 14:01:35
##        today sexe_2 parle
## 1 2021-06-14      1     2
## 2 2021-06-03      1     3
## 3 2021-06-03      1     2
## 4 2021-06-12      1     3
## 5 2021-05-10      0     4
## 6 2021-06-11      1     3
```

Analyses descriptives

La répartion des PME:

```
library(gtsummary) #Pour manipuler les tableaux
#Répartition des chef des PME suivant les variables citées dans le "include"
projet %>% gtsummary::tbl_summary(include = c(sexe, q25, q12, q81))
```

Characteristic	**N = 250**
sexe	
Femme	191 (76%)
Homme	59 (24%)
q25	
Aucun niveau	79 (32%)
Niveau primaire	56 (22%)
Niveau secondaire	74 (30%)
Niveau Superieur	41 (16%)
q12	
Association	6 (2.4%)
GIE	179 (72%)
Informel	38 (15%)
SA	7 (2.8%)
SARL	13 (5.2%)
SUARL	7 (2.8%)
q81	
Locataire	24 (9.6%)
Propriétaire	226 (90%)

```
#On selectionne des variables avec select de dplyr et on les summarise avec la variable sexe, onaffiche
projet %>% dplyr::select(sexe, q12, q25, q81) %>% gtsummary::tbl_summary(by = sexe, percent = "row")
```

Characteristic	**Femme**, N = 191	**Homme**, N = 59
q12		
Association	3 (50%)	3 (50%)
GIE	149 (83%)	30 (17%)
Informel	32 (84%)	6 (16%)
SA	1 (14%)	6 (86%)
SARL	2 (15%)	11 (85%)
SUARL	4 (57%)	3 (43%)
q25		
Aucun niveau	70 (89%)	9 (11%)
Niveau primaire	48 (86%)	8 (14%)
Niveau secondaire	56 (76%)	18 (24%)
Niveau Superieur	17 (41%)	24 (59%)
q81		
Locataire	16 (67%)	8 (33%)
Propriétaire	175 (77%)	51 (23%)

Faisons les statistiques descriptives de notre choix sur les autres variables:

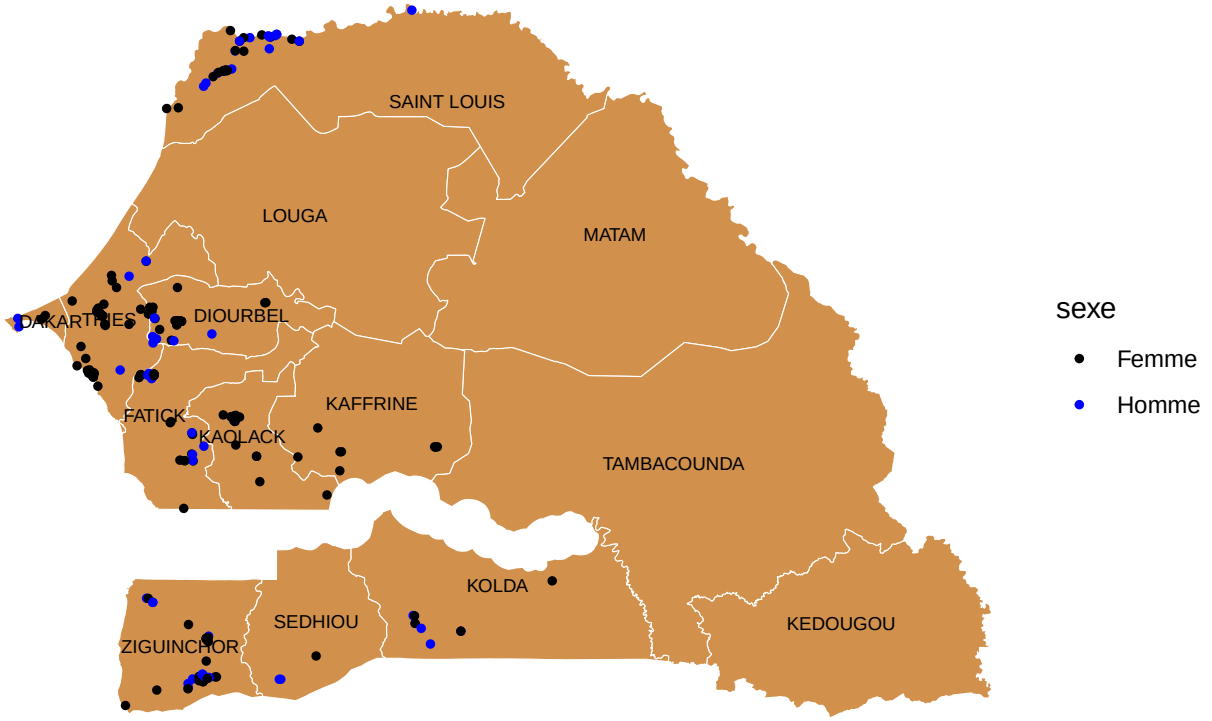
Cartographie

```
library(ggplot2)
library(sf)

projet_map<-st_as_sf(projet,coords= c("gps_menlongitude","gps_menlatitude"),crs=4326)

senegal <- st_read("Limite_Région.shp")
```

```
## Reading layer ‘Limite_RÃ©gion’ from data source
##   ‘C:\Users\DELL\OneDrive\Bureau\Falllou_projet\Limite_RÃ©gion.shp’
##   using driver ‘ESRI Shapefile’
## Simple feature collection with 14 features and 4 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: 227586.3 ymin: 1362012 xmax: 897104.7 ymax: 1845672
## Projected CRS: WGS 84 / UTM zone 28N
```

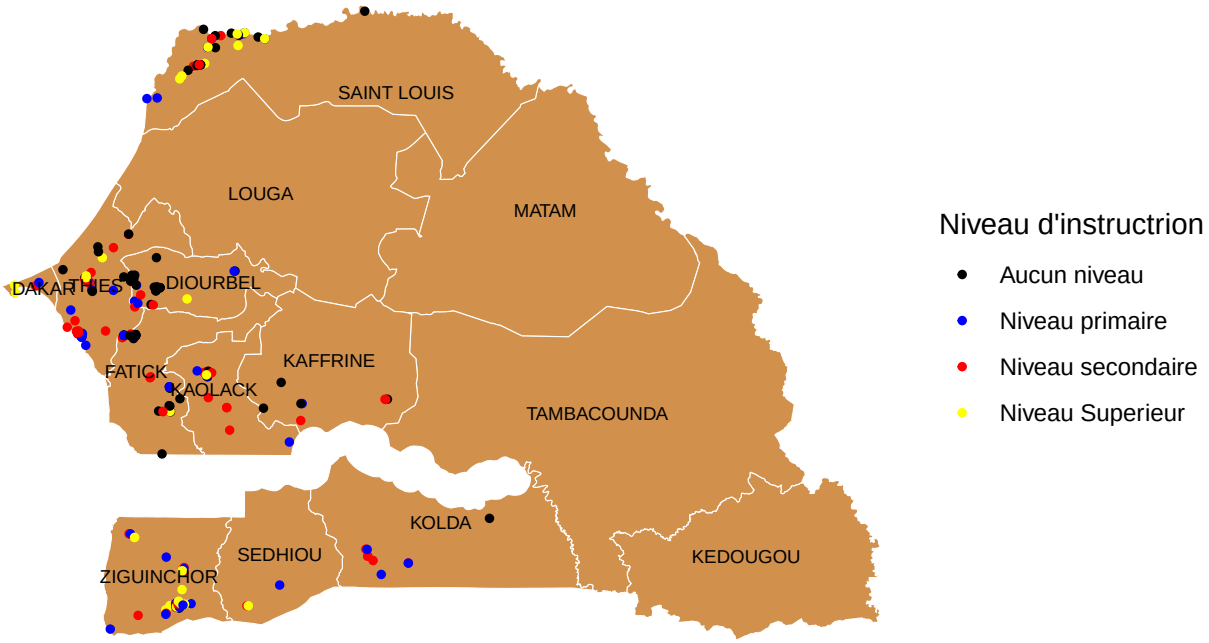
```
names(senegal)[1] <- "region"
ggplot()+
  geom_sf(data=senegal,fill="#D1914D",color="white")+
  geom_sf(data=projet_map,aes(color=sexe),size=1)+
  geom_sf_text(data=senegal,aes(label=region),size=2.5)+
  scale_color_manual(values = c("black", "blue")) +
 theme_void()+
 theme(legend.position = "right")+
 labs(title="Répartion des chefs des PME suivant sexe",color="sexe")
```

Répartion des chefs des PME suivant sexe




```
ggplot()+
  geom_sf(data=senegal,fill="#D1914D",color="white")+
  geom_sf(data=projet_map,aes(color=q25),size=1)+
  geom_sf_text(data=senegal,aes(label=region),size=2.5)+
  scale_color_manual(values = c("black", "blue", "red", "yellow")) +
 theme_void()+
 theme(legend.position = "right")+
 labs(title="Répartion des chefs des PME suivant le Niveau d'instruction",color="Niveau d'instructrion")
```

Répartion des chefs des PME suivant le Niveau d'instruction



Partie 2

• Renommons la variable “country_destination” en “destination” et définissons les valeurs négatives comme manquantes. • Créons une nouvelle variable contenant des tranches d’âge de 5 ans en utilisant la variable “age”. • Créons une nouvelle variable contenant le nombre d’entretiens réalisés par chaque agent recenseur. • Créons une nouvelle variable qui affecte aléatoirement chaque répondant à un groupe de traitement (1) ou de controle (0).

```
#Importation de la fichier Base_Partie 2.xlsx qui est de type excel
#Data se trouve à la feuille 1 du fichier
data <- read_excel("Base_Partie 2.xlsx", sheet = 1)
#District se trouve à la feuille 2 du fichier
district  <- read_excel("Base_Partie 2.xlsx", sheet = 2)
#Codebook se trouve à la feuille 3 du fichier
codebook <- read_excel("Base_Partie 2.xlsx", sheet = 3)
#Renommons les variables comme on avait fait avant
data <- data %>%dplyr::rename(destination = country_destination)
```

```
#classe d'age
cut(data$age, c(15, 20, 25, 30, 35, 40, 45), right = FALSE)
```

```
##  [1] [30,35) [40,45) [25,30) [20,25) [25,30) [20,25) [20,25) [20,25) [20,25)
## [10] [20,25) [20,25) [15,20) [20,25) [15,20) [20,25) [15,20) [15,20) [20,25)
```

```
## [19] [35,40) [15,20) [20,25) [15,20) [25,30) [25,30) [40,45) [15,20) [20,25)
## [28] [40,45) [20,25) [35,40) [30,35) [20,25) [15,20) [30,35) [25,30) [20,25)
## [37] [25,30) [35,40) [20,25) [30,35) [30,35) [25,30) [15,20) [20,25) [25,30)
## [46] <NA>     [15,20) [25,30) [25,30) [40,45) [15,20) [20,25) [35,40) [15,20)
## [55] [20,25) [20,25) [20,25) [25,30) [30,35) [15,20) [20,25) [25,30) [20,25)
## [64] [20,25) [30,35) [25,30) [20,25) [30,35) [20,25) [20,25) [20,25) [20,25)
## [73] [30,35) [25,30) [35,40) [20,25) [30,35) [30,35) [25,30) [25,30) [25,30)
## [82] [35,40) [25,30) [15,20) [20,25) [25,30) [30,35) [25,30) [25,30) [20,25)
## [91] [30,35) [25,30) [15,20) [20,25) [20,25) [15,20) [25,30)
## Levels: [15,20) [20,25) [25,30) [30,35) [35,40) [40,45)
```

```
data[data < 0] <- NA
nb_agents <- data %>%
  group_by(enumerator) %>%
  dplyr::summarise(Nombre_entretiens = n())
nb_agents
```

```
## # A tibble: 16 x 2
##    enumerator Nombre_entretiens
##         <dbl>             <int>
##  1          1                 5
##  2          4                 9
##  3          5                 6
##  4          6                 5
##  5          7                 7
##  6          8                 6
##  7          9                 6
##  8         10                 5
##  9         11                 7
## 10         12                 5
## 11         13                 8
## 12         14                 6
## 13         15                 1
## 14         17                 6
## 15         18                 6
## 16         20                 9
```

```
data$sample <- sample(x = c(0,1), size = length(data$id), replace = TRUE)
data$sample
```

```
##  [1] 0 1 1 1 1 0 0 0 1 1 0 0 1 1 0 0 1 0 0 0 0 0 1 1 1 1 1 1 1 1 0 1 1 0 1 0 1 0
## [39] 1 1 0 0 1 0 1 1 0 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 0 1 0
## [77] 1 0 0 1 0 0 1 0 0 0 1 0 0 1 1 1 1 1 0 0 1
```

```
data$population <- district[data$district, 2]
data$population
```

```
## # A tibble: 97 x 1
##    population
##         <dbl>
##  1      10000
##  2      10000
```

```
##  3      10000
##  4       3000
##  5       3000
##  6      15000
##  7      15000
##  8      15000
##  9      15000
## 10      15000
## # i 87 more rows
```

```
data$duree <- data$endtime-data$starttime
duree_entretien <- data %>%
  group_by(enumerator) %>%
  dplyr::summarise(sum(duree))
```

```
duree_moyen_entretien <- duree_entretien[2] /nb_agents[2]
```

```
for (i in 1:length(colnames(data))){
  colnames(data)[i] <- paste("endline_",colnames(data)[i], sep = "")
}
```

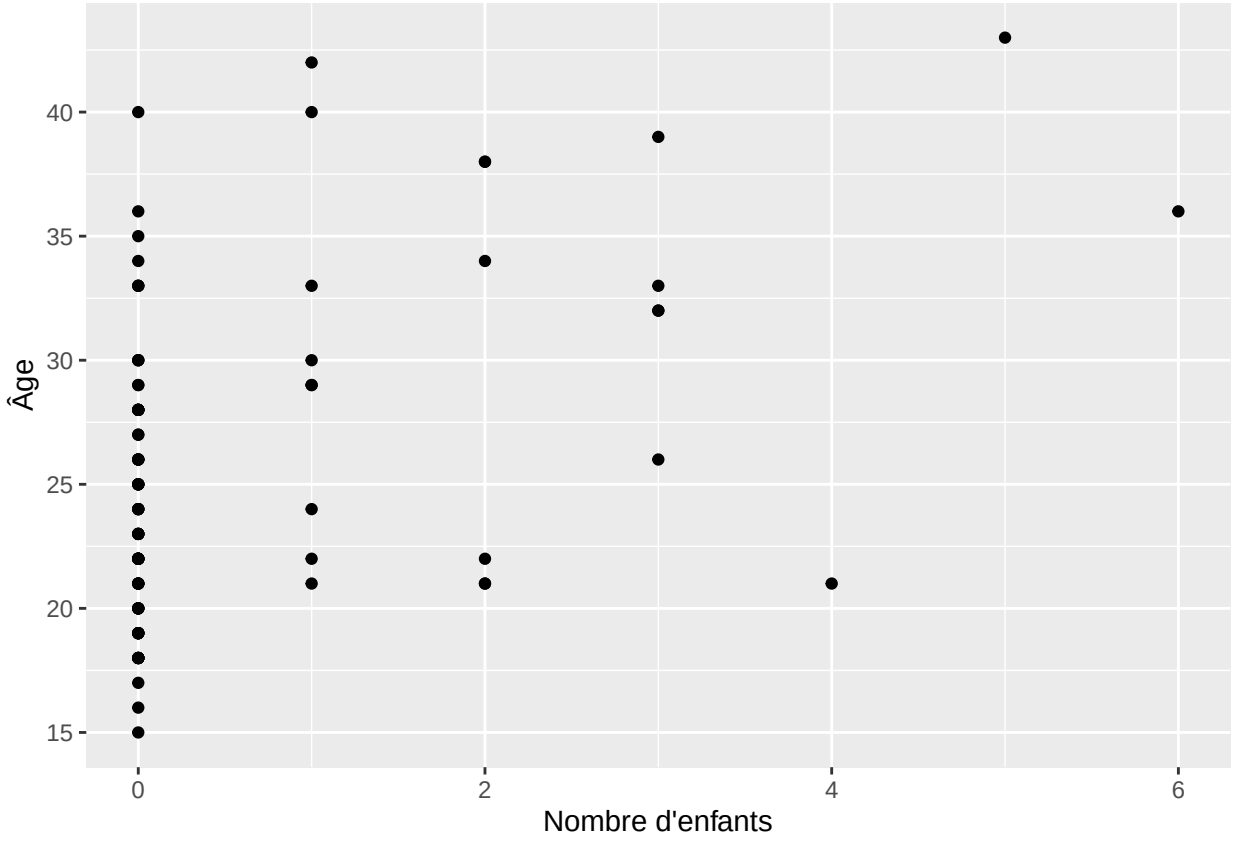
```
data %>%
  group_by(endline_district) %>%
  dplyr::summarise(mean(endline_age), mean(endline_children_num))
```

```
## # A tibble: 8 x 3
##   endline_district `mean(endline_age)` `mean(endline_children_num)`
##              <dbl>               <dbl>                         <dbl>
## 1                1                29.6                         1.5
## 2                2                62.6                         0.852
## 3                3                26.1                         0
## 4                4                26                           0
## 5                5                24.3                         0.5
## 6                6                23.2                         0.115
## 7                7                28                           0.167
## 8                8                24.6                         1.27
```

```
t.test(data$endline_age ~ data$endline_sex, data = data[-46,])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$endline_age by data$endline_sex
## t = -0.95493, df = 10.001, p-value = 0.3621
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -282.7605  113.1009
## sample estimates:
## mean in group 0 mean in group 1
##        25.98837       110.81818
```

```
ggplot(data[-46,]) +
  aes(x = endline_children_num, y = endline_age) +
  geom_point() +
  xlab("Nombre d'enfants") +
  ylab("Âge")
```



SHINY

```
#    http://shiny.rstudio.com/
#
library(sp)
```

```
## Warning: le package 'sp' a été compilé avec la version R 4.1.3
```

```
library(ggplot2)
library(dplyr)
library(shiny)
```

```
## Warning: le package 'shiny' a été compilé avec la version R 4.1.3
```

```
library(leaflet)
```

```
## Warning: le package 'leaflet' a été compilé avec la version R 4.1.3
```

```
library(rnaturalearth)
```

```
## Support for Spatial objects (`sp`) will be deprecated in {rnaturalearth} and will be removed in a fu
```

```
library(rnaturalearthdata)
```

```
## Warning: le package 'rnaturalearthdata' a été compilé avec la version R 4.1.3
```

```
##
## Attachement du package : 'rnaturalearthdata'
```

```
## L'objet suivant est masqué depuis 'package:rnaturalearth':
##
##     countries110
```

```
# Charger les donnÃ©es gÃ©ographiques de l'Afrique de l'Ouest
ne_countries_data <- ne_countries(scale = "medium", continent = "Africa")
```

```
## Warning: The `returnclass` argument of `ne_download()` sp as of rnaturalearth 1.0.0.
## i Please use `sf` objects with {rnaturalearth}, support for Spatial objects
##   (sp) will be removed in a future release of the package.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
west_africa <- subset(ne_countries_data, subregion == "Western Africa")

# Charger les donnÃ©es de base en dehors de la fonction server
base <- read.csv("ACLED-Western_Africa.csv")

ui <- fluidPage(
  # titre de l'application
  titlePanel("shiny map"),

  # Sidebar with a slider input for number of bins
  sidebarLayout(
    sidebarPanel(
      selectInput(
        inputId = "evenement",
        label = "SÃ©lectionnez un evenement",
        choices = c(unique(base$type)),
        selected = "Protests",
        multiple = TRUE

      ),
      selectInput(
```

```
        inputId = "pays",
        label = "SÃ©lectionnez un pays",
        choices = c(unique(base$pays)),
        selected = c(unique(base$pays))[sample(1:length(unique(base$pays)), 1)],
        multiple = TRUE

      ),
      selectInput(
        inputId = "annee",
        label = "SÃ©lectionnez une annee",
        choices = c(unique(base$annee)),
        selected = "2023",
        multiple = TRUE
      ),



    ),
    # Show a plot of the generated distribution
    mainPanel(
      leafletOutput(outputId = "map", width = "100%", height = "720px")
    )
  )
)

server <- function(input, output, session) {
  filtered_data <- reactive({
    subset(base, pays %in% input$pays & type %in% input$evenement & annee %in% input$annee)
  })

  output$map <- renderLeaflet({
    filtered_west_africa <- west_africa[west_africa$name %in% input$pays]

    leaflet() %>%
      addProviderTiles(providers$Stamen.Toner) %>%
      addPolygons(data = ne_countries(type = "countries", country = input$pays), fillColor = "lightblue


      addCircleMarkers(data = filtered_data(),
                       lat = ~latitude,
                       lng = ~longitude,
                       radius = 3,
                       opacity = 0.7)
  })
}

shinyApp(ui = ui, server = server)
```

```
##
## Listening on http://127.0.0.1:5814
```

shiny map

SÃ©lectionnez un evenement

Protests

SÃ©lectionnez un pays

Guinea

SÃ©lectionnez une annee

2023