Rendu rapport du projet

Nicolas\_LOUGUE

2023-07-24

# Partie 1

## Importation et mise en forme

Nous chargeons tous les packages que nous aurons besoin pour notre travail.

Après avoir charger tous les packages que nous aurons besoin au cours de notre TP, nous importons la base d’étude

# importation de la base de données  
projet <-data.frame(readxl:: read\_xlsx("Base\_Partie 1.xlsx"))

Tableau des valeurs manquantes par variables

# Calculer le nombre de valeurs manquantes par variable  
NA\_tableau <- projet %>%  
 summarise(across(everything(), ~ sum(is.na(.)))) %>%  
 pivot\_longer(everything(), names\_to = "variable", values\_to = "Nombre de valeurs manquantes")  
  
Tableau <- gt(NA\_tableau, rowname\_col = "rowname")  
Tableau

| variable | Nombre de valeurs manquantes |
| --- | --- |
| key | 0 |
| q1 | 0 |
| q2 | 0 |
| q23 | 0 |
| q24 | 0 |
| q24a\_1 | 0 |
| q24a\_2 | 0 |
| q24a\_3 | 0 |
| q24a\_4 | 0 |
| q24a\_5 | 0 |
| q24a\_6 | 0 |
| q24a\_7 | 0 |
| q24a\_9 | 0 |
| q24a\_10 | 0 |
| q25 | 0 |
| q26 | 0 |
| q12 | 0 |
| q14b | 1 |
| q16 | 1 |
| q17 | 131 |
| q19 | 120 |
| q20 | 0 |
| filiere\_1 | 0 |
| filiere\_2 | 0 |
| filiere\_3 | 0 |
| filiere\_4 | 0 |
| q8 | 0 |
| q81 | 0 |
| gps\_menlatitude | 0 |
| gps\_menlongitude | 0 |
| submissiondate | 0 |
| start | 0 |
| today | 0 |

Vérification si la clé key ne contient pas de valeurs manquantes

# Vérifier si la variable "key" contient des valeurs manquantes  
  
ifelse(any(is.na(projet$key)),"valeurs manquantes","La variable key n'a pas de valeurs manquantes")

## [1] "La variable key n'a pas de valeurs manquantes"

## Création de variables

Renommage et création de nouvelles variables Là, nous renommons respectivement les variables “q1”, “q2” et “q23” en region, departement et sexe.

projet\_1 <- projet %>% rename(region=q1, departement=q2, sexe= q23)  
projet\_1 <- projet\_1 %>%  
 dplyr::mutate(sexe\_2=recode(sexe,"Femme"="1","Homme"="0"))  
  
langues <- data.frame(key=projet\_1$key, dplyr::select(projet\_1,starts\_with("q24a")))  
langues <- langues %>% dplyr::mutate(parle=rowSums(langues[,2:length(langues)]))  
langues <- langues %>% dplyr::select(key,parle)  
data <- merge.data.frame(projet\_1, langues, by= "key")

## Analyse descriptive

### Analyse descriptive univariée

##### Types de variables  
stat\_des <- data[c("sexe", "q25","q12","q81")]  
stat\_des\_1 <- stat\_des %>% rename( "Niveau instruction"=q25, "Statut juridique"=q12, "Propriétaire/Non propriétaire"=q81)

Résultats

# Summary  
stat\_des\_1 %>% tbl\_summary() #table de résumé des variables d'intérêt

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **N = 250**1 |
| --- | --- |
| sexe |  |
| Femme | 191 (76%) |
| Homme | 59 (24%) |
| Niveau instruction |  |
| Aucun niveau | 79 (32%) |
| Niveau primaire | 56 (22%) |
| Niveau secondaire | 74 (30%) |
| Niveau Superieur | 41 (16%) |
| Statut juridique |  |
| Association | 6 (2.4%) |
| GIE | 179 (72%) |
| Informel | 38 (15%) |
| SA | 7 (2.8%) |
| SARL | 13 (5.2%) |
| SUARL | 7 (2.8%) |
| Propriétaire/Non propriétaire |  |
| Locataire | 24 (9.6%) |
| Propriétaire | 226 (90%) |
| 1n (%) | |

### Analyse descriptive bivariée

tc\_1 <- stat\_des\_1 %>%  
 gtsummary::tbl\_cross(  
 row = "Statut juridique",  
 col = sexe,  
 percent = "row"  
 ) %>%  
 add\_p(source\_note = FALSE) ## ajouter les p-value en précisant le test réalisé  
  
tc\_2 <- stat\_des\_1 %>%  
 gtsummary::tbl\_cross(  
 row = "Niveau instruction",  
 col = sexe,  
 percent = "row"  
 ) %>%  
 add\_p(source\_note = FALSE) ## ajouter les p-value en précisant le test réalisé  
  
tc\_3 <- stat\_des\_1 %>%  
 gtsummary::tbl\_cross(  
 row = "Propriétaire/Non propriétaire",  
 col = sexe,  
 percent = "row"  
 ) %>%  
 add\_p(source\_note = FALSE) ## ajouter les p-value en précisant le test réalisé

Résultats tableau de contingence

gtsummary::tbl\_stack(tbls=list(tc\_1, tc\_2, tc\_3))

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

|  | sexe | |  | |
| --- | --- | --- | --- | --- |
|  | Femme | Homme | Total | p-value1 |
| Statut juridique |  |  |  | <0.001 |
| Association | 3 (50%) | 3 (50%) | 6 (100%) |  |
| GIE | 149 (83%) | 30 (17%) | 179 (100%) |  |
| Informel | 32 (84%) | 6 (16%) | 38 (100%) |  |
| SA | 1 (14%) | 6 (86%) | 7 (100%) |  |
| SARL | 2 (15%) | 11 (85%) | 13 (100%) |  |
| SUARL | 4 (57%) | 3 (43%) | 7 (100%) |  |
| Total | 191 (76%) | 59 (24%) | 250 (100%) |  |
| Niveau instruction |  |  |  | <0.001 |
| Aucun niveau | 70 (89%) | 9 (11%) | 79 (100%) |  |
| Niveau primaire | 48 (86%) | 8 (14%) | 56 (100%) |  |
| Niveau secondaire | 56 (76%) | 18 (24%) | 74 (100%) |  |
| Niveau Superieur | 17 (41%) | 24 (59%) | 41 (100%) |  |
| Total | 191 (76%) | 59 (24%) | 250 (100%) |  |
| Propriétaire/Non propriétaire |  |  |  | 0.2 |
| Locataire | 16 (67%) | 8 (33%) | 24 (100%) |  |
| Propriétaire | 175 (77%) | 51 (23%) | 226 (100%) |  |
| Total | 191 (76%) | 59 (24%) | 250 (100%) |  |
| 1Fisher's exact test | | | | |

## Cartographie

Nous chargeons les packages utiles pour faire de la cartographie

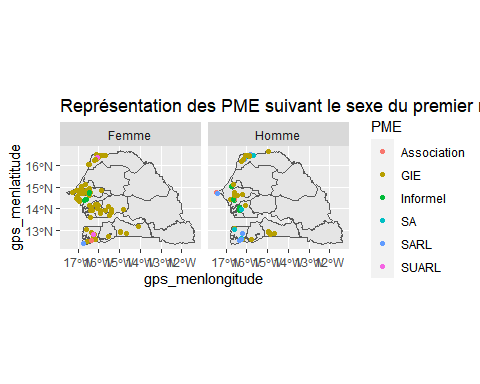
library(sf)  
library(rnaturalearth)  
library(raster)  
library(RColorBrewer)

Conversion du tableau data frame en données géographiques

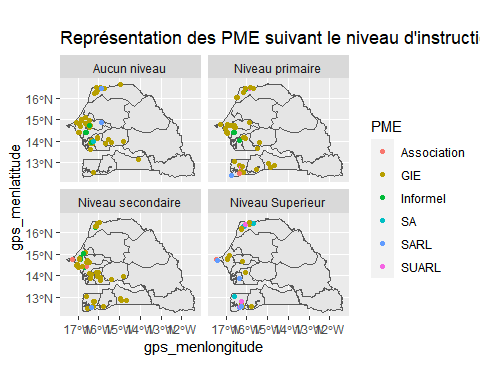
projet\_map <- projet\_1 %>% rename(PME=q12)

Représentation spatiale

senegal\_1 <- ne\_states(country = "senegal", returnclass = "sf")  
   
ggplot() +  
 geom\_sf(data = senegal\_1) +  
 geom\_point(data = projet\_map, aes( x =gps\_menlongitude , y = gps\_menlatitude,color = PME)) +  
 facet\_wrap(~sexe)+  
 labs(title = "Représentation des PME suivant le sexe du premier responsable")



ggplot() +  
 geom\_sf(data = senegal\_1) +  
 geom\_point(data = projet\_map, aes( x =gps\_menlongitude , y = gps\_menlatitude,color = PME)) +  
 facet\_wrap(~q25)+  
 labs(title = "Représentation des PME suivant le niveau d'instruction")



# Partie 2

## Nettoyage de la base

data <- readxl::read\_xlsx("Base\_Partie 2.xlsx", sheet=1)  
  
data\_1 <- data%>%  
 dplyr::mutate(destination=ifelse(country\_destination < 0, NA,country\_destination))  
  
  
# Création de la variable de tranche de 5 ans  
data\_1 <- data\_1%>%  
 dplyr:: mutate(age\_intervalle =cut(age,include.lowest = TRUE,  
 right = TRUE, dig.lab = 9,  
 breaks = c(15, 20, 25, 30, 35, 40, 45, 50, 55, 65)))  
  
   
  
set.seed(123)  
data\_2 <- data\_1%>%  
 dplyr:: mutate(groupe=sample(c(0,1), nrow(data\_1), replace=TRUE))  
  
  
tab=as.data.frame(table(data\_2$enumerator))  
tab<- tab%>%  
 dplyr:: rename(enumerator=Var1, "nombre d'entretien"=Freq)  
  
data\_3<- merge(data\_2,tab,by="enumerator", all=TRUE)  
  
  
data\_4<- readxl::read\_xlsx("Base\_Partie 2.xlsx", sheet=2)  
  
data\_5<- merge(data\_3,data\_4, by="district",all=TRUE)  
  
data\_6 <- data\_5 %>% dplyr::mutate(duree\_entretien=endtime-starttime)   
  
duree\_moyenne <- dplyr::select(data\_6, c("enumerator", "duree\_entretien")) %>%  
 dplyr::rename(Id\_enquêteur=enumerator) %>%  
 group\_by(Id\_enquêteur) %>%   
 summarize(duree\_moyen=mean(duree\_entretien))   
   
   
duree\_moyenne %>% flextable()

| Id\_enquêteur | duree\_moyen |
| --- | --- |
| 1 | 68.14667 mins |
| 4 | 36.48333 mins |
| 5 | 33.55833 mins |
| 6 | 25.84667 mins |
| 7 | 37.16429 mins |
| 8 | 40.13056 mins |
| 9 | 114.76667 mins |
| 10 | 55.27667 mins |
| 11 | 33.48333 mins |
| 12 | 48.16667 mins |
| 13 | 31.59583 mins |
| 14 | 25.56111 mins |
| 15 | 28.65000 mins |
| 17 | 29.28611 mins |
| 18 | 36.85833 mins |
| 20 | 28.76852 mins |

# Renommer les variables d'ensemble avec le préfixe "endline"  
  
noms\_colonnes =apply(matrix(colnames(data\_6)), 1, function(nom\_colonne) {  
 paste0("endline\_", nom\_colonne)  
})  
  
data\_7 <- data\_6  
names(data\_7) <- unlist(noms\_colonnes)

## Analyse et visualisation des données

Tableau recapitulatif contenant l’âge moyen et le nombre d’enfants moyen par district

tab\_recap <- data\_7 %>%  
 group\_by(endline\_district) %>%  
 summarize(age\_moyen = mean(endline\_age),  
 enfants\_moyen = mean(endline\_children\_num))  
tab\_recap %>% flextable()

| endline\_district | age\_moyen | enfants\_moyen |
| --- | --- | --- |
| 1 | 29.62500 | 1.5000000 |
| 2 | 62.62963 | 0.8518519 |
| 3 | 26.12500 | 0.0000000 |
| 4 | 26.00000 | 0.0000000 |
| 5 | 24.33333 | 0.5000000 |
| 6 | 23.15385 | 0.1153846 |
| 7 | 28.00000 | 0.1666667 |
| 8 | 24.63636 | 1.2727273 |

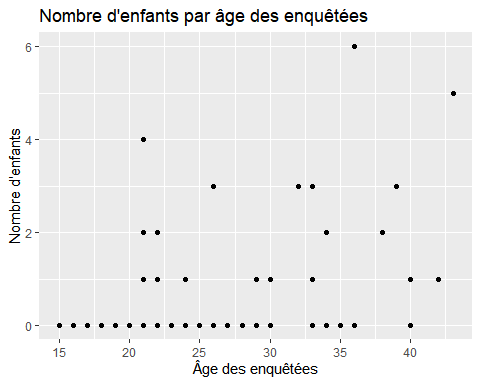
Test de différence

data\_7$Female<-ifelse(data\_7$endline\_sex==0,1,0)  
data\_7$Male<-ifelse(data\_7$endline\_sex==1,1,0)  
tableau\_recap <- data\_7 %>%  
 tbl\_summary(include=c(Female,Male,endline\_age),by = endline\_sex, statistic = list(Male~"{p}%", Female~"{p}%",endline\_age ~ "{mean}")) %>%  
 #add\_p(test = endline\_age~ "t.test") %>%  
 add\_difference(test=list( Female ~ "t.test", Male ~ "t.test"), pvalue\_fun=scales::label\_pvalue(accuracy = .0001)  
)  
tableau\_recap

| **Characteristic** | **0**, N = 861 | **1**, N = 111 | **p-value**2 | **Difference**2 | **95% CI**23 |
| --- | --- | --- | --- | --- | --- |
| Female | 100% | 0% |  |  |  |
| Male | 0% | 100% |  |  |  |
| endline\_age | 26 | 111 | 0.3621 | -85 | -283, 113 |
| 1%; Mean | | | | | |
| 2Welch Two Sample t-test | | | | | |
| 3CI = Confidence Interval | | | | | |

Nuage de point

nuage\_point <- ggplot(filter(data\_7, !(endline\_age == 999)), aes(x = endline\_age, y = endline\_children\_num)) + geom\_point() +labs(x = "Âge des enquêtées", y = "Nombre d'enfants", title = "Nombre d'enfants par âge des enquêtées")  
nuage\_point



Estimation de l’effet d’appartenance à un groupe

modele <-stats::lm(data\_7$endline\_intention ~ data\_7$endline\_groupe , data = data\_7)  
modele%>% gtsummary::tbl\_regression()

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **Beta** | **95% CI**1 | **p-value** |
| --- | --- | --- | --- |
| data\_7$endline\_groupe | 0.34 | -0.36, 1.0 | 0.3 |
| 1CI = Confidence Interval | | | |

Tableau avec 3 modèle

modele\_A <- modele%>% gtsummary::tbl\_regression()  
modele\_B<- lm(data\_7$endline\_intention ~ data\_7$endline\_age+data\_7$endline\_sex, data = data\_7)%>%tbl\_regression()  
modele\_C<- lm(data\_7$endline\_intention ~ data\_7$endline\_age + data\_7$endline\_sex + data\_7$endline\_district, data = data\_7)%>%tbl\_regression()  
  
gtsummary::tbl\_stack(  
list(modele\_A,modele\_B,modele\_C),  
## intitulé des groupes de tableau associés  
group\_header = c("Modèle A","Modèle B","Modèle C")  
)

| **Group** | **Characteristic** | **Beta** | **95% CI**1 | **p-value** |
| --- | --- | --- | --- | --- |
| Modèle A | data\_7$endline\_groupe | 0.34 | -0.36, 1.0 | 0.3 |
| Modèle B | data\_7$endline\_age | 0.00 | 0.00, 0.00 | 0.9 |
|  | data\_7$endline\_sex | -0.90 | -2.0, 0.23 | 0.12 |
| Modèle C | data\_7$endline\_age | 0.00 | 0.00, 0.00 | >0.9 |
|  | data\_7$endline\_sex | -0.85 | -2.0, 0.29 | 0.14 |
|  | data\_7$endline\_district | 0.09 | -0.07, 0.24 | 0.3 |
| 1CI = Confidence Interval | | | | |

# Application shiny

Cette partie est dédiée à la mise en place de l’application shiny concernant le présent projet. Vous trouverez dans le document ci-joint le fichier app.R dans le dossier shiny\_app\_projet\_cartographie.