République du Sénégal

Un peuple – Un but – Une foi

****

Ministère de l'Économie du Plan et de la Coopération

****

Agence Nationale de la statistique et de la Démographie

(ANSD)

École Nationale de la Statistique et de l'Analyse Économique Pierre Ndiaye

(ENSAE)

Cours : Traitement et analyse des données sur R

Projet final

Rédigé par :

Brahima TOU

Elève Ingénieur Statisticien Economiste en 3e année

Sous la supervision de

M. HEMA

Research-Analyst

Juillet 2023

Sommaire

List of Figures

List of Tables

Résumé

Dans cet excercice très partique,il etait question de mettre en pratique les connaissances apprises durant le ,module intitulé rapport statistique sur R.Le travail est contituié de trois grandes parties.La partie 1 qui traite essentiellement de l'utilisation du package gtsummary pour les sorties des analyses descriptes et de la cartographie et lesdonnées utilisées renseignent sur les PME du Sénégal.Il ressort de cette partie qu'on peut fusionner des tabbleaux de statistiques descriptives pour un nombre de variables depassant meme 2.De meme, la partie cartographie nous a familiarisé avec la repartition spatiale d'une variable dans l'espace d'un pays.La deuxième partie concerne la manipulation des fonction sur R pour traiter, apurer, creer des variables afin d'etre plus souple et concis dans les analyses.Pour faire cela, une base artificelle a été mise à notre disposition dans le cadre de l'etude.Concernant la troixième partie de l'étude, elle consiste à la creation d'une application shiny permettant de visualiser les informations d'une base.A cet effet, base ACLED-Western_Africa.csv decrivanta les evenements (violences politiques, marches, ...) en Afrique de l'Ouest a été mise à notre disposition pour des fins de l'étude.

Introduction

Développé par Ross Ihaka et Robert Gentleman,R est un langage de programmation pour l'analyse et la modélisation des données. R peut être utilise comme un langage oriente objet tout comme un environnement statistique dans lequel des listes d'instructions peuvent être exécutées en séquence sans l'intervention de l'utilisateur.De part sa puissance dans le traitement des données, R s'est imposé dans le quotidien des statisticiens et est devenu un logiciel incontournable chez les statisticiens.C'est ainsi que les élèves ISE cycle long et les ISE première année ont dans leurs programmes, le cours de Projet Statistique avec R.Durant les 20h du cours, nous avaons pris à traiter , analyser et meme à creer nos propres fonctions sur R et ceici a été plus approfondi avec des exposés riches de nouvelles idées.Une fois ces connaissances acquises, l'heure est venue pour les mettre en pratique dans l'optique de les consolider.C'est dans cette optique ce projet s'inscrit pour nous permettre d'approfondir nos connaissances sur le logiciel sur R mais des boien consolider les acquis.

Pour bien mener notre travail suivra l'ordre des parties ainsi que des questions.Dans la première partie, nous aborderons toutes les questions dans l'ordre.Ensuite, nous ferons la deuxième partie dans l'ordre des questions et enfin on terminera par latroixème partie.

1.1. Préparation des données

1.1.1. Importation et mise en forme

```
#install.packages("haven",dependencies=TRUE)
library(haven)
library(readxl)
##Base_Partie1 <- read_excel(paste0(chemin,"\\Base_Partie 1.xlsx"))

Base_Partie1 <- read_excel("Base_Partie 1.xlsx")##importation de la base de format excel

projet<-data.frame(Base_Partie1) ##convertion de la base en un objet data.frame
```

1.1.1.1. Importation de la base de l'étude

1.1.1.2. Faites un tableau qui resume les valeurs manquantes par variable Nous calculerons aussi les pourcentages des valeurs manquantes de chaque variable par rapport au nombre total de valeurs manquantes de la base.

```
#calcul des pourcentages des valeurs manquantes
percen=colSums((is.na(projet)))*100/nrow(projet) #calcul des effectifs des valeurs manquantes
effectifs=(colSums(is.na(projet))) ##regroupement des elements dans un data.frame
```

Table 1: Résumé des valeurs manquantes

	effectifs	pourcentages
key	0	0.0
q1	0	0.0
q2	0	0.0
q23	0	0.0
q24	0	0.0
q24a_1	0	0.0
q24a_2	0	0.0
q24a_3	0	0.0
q24a_4	0	0.0
q24a_5	0	0.0
q24a_6	0	0.0
q24a_7	0	0.0
q24a_9	0	0.0
q24a_10	0	0.0
q25	0	0.0
q26	0	0.0
q12	0	0.0
q14b	1	0.4
q16	1	0.4
q17	131	52.4
q19	120	48.0
q20	0	0.0
filiere_1	0	0.0
filiere_2	0	0.0
filiere_3	0	0.0
filiere_4	0	0.0
q8	0	0.0
q81	0	0.0
gps_menlatitude	0	0.0
gps_menlongitude	0	0.0
submissiondate	0	0.0
start	0	0.0
today	0	0.0

```
manques<-data.frame(effectifs=effectifs,pourcentages=percen) ## affichage du resultat des valeurs manqu
manques %>% kable(format = "latex",caption = " Résumé des valeurs manquantes")
```

1.1.1.3. Vérifier s'il y a des valeurs manquantes pour la variable key dans la base projet

On note qu'il n'existe pas de valeurs manquantes dans la variable Key.On remarque qu'il n'existe pas de valeurs manquantes.

```
attach(projet)
which(is.na(key),arr.ind=TRUE)
```

```
## integer(0)
```

1.1.2. Création de variables

On renomme facilement l'ensemble des variiables puis on utilise la **fonction names()**pour s'assurer que les noms ont été bien changés

```
##Renomme chaque variable derriere l'églité au nom avant l'égalité
projet<-projet %>%
  rename(region=q1, departement=q2, sexe=q23)
head(projet)[,2:4]%>%
  kable(format = "latex")
```

region	departement	sexe
Diourbel	Bambey	Femme
Thiès	Mbour	Femme
Thiès	Mbour	Femme
Thiès	Mbour	Femme
Ziguinchor	Bignona	Homme
Ziguinchor	Oussouye	Femme

1.1.2.1. Créer la variable sexe_2 qui vaut 1 si sexe égale à Femme et 0 sinon

Il s'agit d'un recodage de la variable sexe .On utilise la fonction **ifelse()** qui testera si la condition, met la valeur correcte femme et 0 sinon.

```
projet$sexe_2<-ifelse(projet$sexe=="Femme",1,0)##la fonction ifelse verifie si sexe est égale femme, il
```

1.1.2.2. Créer un data.frame nommé langues

Ce data.frame prend les variables **key** et les variables correspondantes décrites plus haut.

```
langues<-data.frame(projet %>%
  select("key","q24a_1","q24a_2","q24a_3","q24a_4","q24a_5","q24a_6","q24a_7","q24a_9","q24a_10"))##sel
```

1.1.2.3. Créer une variable parle

Cette variable est égale au nombre de **langue** parlée par le dirigeant de la PME.

```
langues$parle<-rowSums(langues[,2:10])##extraction de toutes les lignes et descolonnes  2 jusqu'à 10
head(langues)["parle"]%>%
  kable(format = "latex")
```

parle
2
3
2
3
4
3

```
langues<-data.frame(langues %>%
  select("key","parle"))
head(langues)%>%
  kable(format = "latex")
```

1.1.2.4. Sélectionnez uniquement les variables key et parle, l'objet de retour sera langues.

key	parle
uuid:68bff42b-1228-4c66-9bcc-e6d312d9fea6	2
uuid:d70b3c7e-3ca0-4358-bc59-3f7f6baf55e9	3
uuid:0ac18b64-7d85-4bb9-a842-698ac79909af	2
uuid:c52cf5e4-8c28-4e65-998b-3fe2a971a1a3	3
uuid:ac177870-001c-4ada-8747-c22ffe4e4596	4
uuid:578097cf-9af7-46e6-8992-d9079b14c342	3

1.1.2.5. Merger les data.frame projet et langues

Danc ce cas,on utilise la fonction merge avec comme clé d'identification la variable Key car elle est la variable commmune aux deux bases et c'est elle qui permet d'identifier les individus.

```
projet_parle<-data.frame(projet %>%
                           merge(langues,by="key"))## la fonction merge dans notre cas fusionne la base
dim(projet_parle)## verifier si le merge a été effectué avec succès
```

```
## [1] 250  35
```

1.2. Analyses descriptives

1.2.1. Analyse de la repartition des PM par filière suivant le sexe, statut juridique, propritaire

Il resssort du tableauci-dessous que les PME sont dirigés majoritairement par des femmes quel qu'en soit la filière.Concernant le titreproprietaire ou locataire, on note une grande predominance des propritaire(plus de 4 sur 5 ont le statut de proprietaire) quel qu'en soit la filière concernée.La repartition suvant le niveau d'instruction du responsable montre que moins de 50 % de ces responsables de PME ont le niveau secondaire et cela pour les filières **arcachide, anacarde, mangues**.Quant àla filière **riz**, plus de 50% ont au moins le niveau superieur.

En s'interressant au statut juridique, on remarque une grande predominance des GIE (plus de 70% des PME) sont des GIE toutes les filières confondues.De facon inverse, dans toutes les filières, on note une faible presence des PME ayant le statut **SA,SAR,SUARL** .

```
## Supreesion  les espaces

theme_gtsummary_compact(set_theme = TRUE, font_size = NULL)
## Format de la sortie

theme_gtsummary_printer(
  print_engine = "flextable",  #c("gt", "kable", "kable_extra", "flextable",
  #"huxtable", "tibble"),
  set_theme = TRUE
)

## CRÉONS UNE FONCTION QUI NOUS FACILITE LA CRÉATION DES TABLEAU

tabl_filiere = function(base_donnee, num_var_filiere_, nom_filiere){

  ### TITRE DE LA COLONNE

  nom <- paste(paste("", as.character(nom_filiere), sep = ""), "", sep = "")

  ### TABLEAU GTSUMMARY
```

```
  tableau <- base_donnee %>%
    dplyr::select(sexe_2, sexe, q25, q12, q81, names(base_donnee[num_var_filiere_])
    ) %>%
    gtsummary::tbl_summary(
      ## paramètres de tbl_summary
      include = c(names(base_donnee[num_var_filiere_]), sexe_2,q81, q25, q12),
      by = names(base_donnee[num_var_filiere_]),## variables qui forme les groupes
      label = list(sexe_2 ~ "Responsable femme",
                   q25 ~ "Niveau d'instruction du responsable",
                   q12 ~ "statut juridique de l'entreprise",
                   q81 ~ "Propriétaire / locataire"
      ), ## labélisation des variables dans le tableau
      percent = "column", ## Type de pourcentage affichés dans le tableau
      digits = ~2, ## nombre de chiffre après la virgule pour les résultats
      statistic = c(all_categorical(), all_interaction()) ~ "{p}% ({n})",
      type = list(sexe_2 ~ "dichotomous"), ## modifier et préciser comment il
      ## faut considérer la variable SEXE_2
      missing = "ifany", ## afficher les stat sur les valeurs manquantes
      missing_text = "Manquants", ## formatage et nomination de "valeur manquante"
    ) %>%
    ## ajouter les statistiques sur la base totale (non par groupe)
    add_overall() %>%
    ## mise en forme des variables et des modalités
    bold_labels() %>%
    italicize_levels() %>%
    ## mise en forme de l'entête du tableau
    modify_header(
      list(
        label ~ "Variable",
        stat_0 = "TOTAL (n={N})",
        all_stat_cols(stat_0 = FALSE) ~ "{level} (n={n}, {style_percent(p)}%)",
        stat_2 = paste(nom,", (n={n}, {style_percent(p)}%)", sep = "")
      )
    ) %>%
    modify_column_hide(c(stat_0,stat_1))
  return(tableau)
}

# STAT. POUR LA FILIÈRE ARACHIDE
which(names(projet) == "filiere_1")
```

```
## [1] 23
```

```
tabl_fil_ara <- tabl_filiere(projet, 23, "Arachide")
# STAT. POUR LA FILIÈRE ANACARDE
which(names(projet) == "filiere_2")
```

```
## [1] 24
```

```
tabl_fil_ana <- tabl_filiere(projet, 24, "Anacarde")
# STAT. POUR LA FILIÈRE MANGUE
which(names(projet) == "filiere_3")
```

```
## [1] 25
```

```r
tabl_fil_man <- tabl_filiere(projet, 25, "Mangue")
# STAT. POUR LA FILIÈRE RIZ
which(names(projet) == "filiere_4")
```

```
## [1] 26
```

```r
tabl_fil_riz <- tabl_filiere(projet, 26, "Riz")

# tableau aggrégé
tabl_filiere_ <- gtsummary::tbl_merge(
  list(tabl_fil_ara, tabl_fil_ana, tabl_fil_man, tabl_fil_riz), tab_spanner =  c("Arachide", "Anarchad

tabl_filiere_ %>% modify_caption("Repartition des PM produisant la mangue et le riz en fonction sexe,n
```

Table 2: Repartition des PM produisant la mangue et le riz en fonction sexe,niveau d'instruction, Proprietaire et du statut juridique

	Arachide	Anarchade	Mangue	Riz
Variable	Arachide, (n=108, 43%)[1]	Anacarde, (n=61, 24%)[1]	Mangue, (n=89, 36%)[1]	Riz, (n=92, 37%)[1]
Responsable femme	86.11% (93.00)	65.57% (40.00)	76.40% (68.00)	83.70% (77.00)
Propriétaire / locataire				
Locataire	11.11% (12.00)	11.48% (7.00)	12.36% (11.00)	9.78% (9.00)
Propriétaire	88.89% (96.00)	88.52% (54.00)	87.64% (78.00)	90.22% (83.00)
Niveau d'instruction du responsable				
Aucun niveau	39.81% (43.00)	21.31% (13.00)	29.21% (26.00)	11.96% (11.00)
Niveau primaire	21.30% (23.00)	27.87% (17.00)	26.97% (24.00)	28.26% (26.00)
Niveau secondaire	31.48% (34.00)	24.59% (15.00)	28.09% (25.00)	34.78% (32.00)
Niveau Superieur	7.41% (8.00)	26.23% (16.00)	15.73% (14.00)	25.00% (23.00)
statut juridique de l'entreprise				
Association	1.85% (2.00)	4.92% (3.00)	0.00% (0.00)	2.17% (2.00)
GIE	73.15% (79.00)	57.38% (35.00)	82.02% (73.00)	83.70% (77.00)
Informel	21.30% (23.00)	19.67% (12.00)	5.62% (5.00)	3.26% (3.00)
SA	1.85% (2.00)	3.28% (2.00)	3.37% (3.00)	3.26% (3.00)
SARL	0.93% (1.00)	9.84% (6.00)	6.74% (6.00)	5.43% (5.00)
SUARL	0.93% (1.00)	4.92% (3.00)	2.25% (2.00)	2.17% (2.00)

[1]% (n)

1.2.2. Répartition filière suivant le croisement de deux variables

L'analyse repartition des filières suivant le sexe revèle que 76% de PME de Mangues appartiennent aux femmes contre 24% chez les hommes.Les memes tendances sont observées au niveau de la filière riz ou 77,84% des PME ont pour proprietaire femme.En fonction du niveau d'instruction, les femmes semblent avoir le niveau d'instruction le plus faible qu'il soit de la filière mangue ou riz.

Suivant le statut jurique, 62% des PME de la filière mangues sont des GIE et appartiennent aux femmmes contre 12% des GIE des hommes.Les memes ramarques sont faites au niveau de la filière riz avec comme seule particularuité 2% des PME des hommes qui sont des associations.En ce qui concerne si le proˆpriataire est femme ou homme, la p-value est superieur à 5% donc nous nous reservons de les commenter.Les memes remarques sont faites au niveau de la repartition en fonction du sexe dans lafilière arachide et anacarde.Les seules exceptions avec les deux precedentes resultent au fait que la p-value du niveau d'instruction pour les femmes n'est pas significative doncinutile de commenter ces chiffres.

```
theme_gtsummary_compact(set_theme = TRUE, font_size = NULL)
#Créons une fonction pour les tableaux croisés par filière et sexe

tabl_fil =function(base_donnee, num_var_filiere_, nom_filiere,lab_var,num_var){

  ### TABLEAU CROISE

  tableau <- base_donnee %>%
    dplyr::select(sexe, q25, q12, q81, names(base_donnee[num_var_filiere_])
    ) %>%
    gtsummary::tbl_strata(
      strata = names(base_donnee[num_var_filiere_]),
      .tbl_fun = ~ .x %>%
        gtsummary::tbl_cross(
          row = names(base_donnee[num_var]),
          col = sexe,
          percent = "cell",
          margin = NULL,
          #statistic = ~ "{p}% ({n})",
          #digits = ~ 2,
          label = list(names(base_donnee[num_var]) ~ as.character(lab_var),
                       sexe ~ "sexe du responsable")
        ) %>% add_p() %>%
        bold_labels() %>%
        italicize_levels(),

      ## préciser comment combiner les tableaux de chaque groupe. Par défaut,
      ## il combine avec "tbl_merge"
      .combine_with = "tbl_merge",
      .header = "{strata}"
    )%>%
    ## mise en forme de l'entête du tableau
    modify_header(
      list(
        all_stat_cols(stat_0 = FALSE) ~ "**{level}** (n={n}, {style_percent(p)}%)"
      )
    ) %>%
    modify_column_hide(c(stat_1_1,stat_2_1,p.value_1)) #%>% modify_spanning_header(variable ~ paste(nom
  return(tableau)
}

tab_prop_1 <- tabl_fil(projet,23,"Arachide","Propriétaire/Locataire",which(names(projet)=="q81"))
tab_niv_1 <- tabl_fil(projet,23,"Arachide","Niveau d'instruction",which(names(projet)=="q25"))
tab_stat_1 <- tabl_fil(projet,23,"Arachide","Statut juridique",which(names(projet)=="q12"))

tab_fil_1 <- gtsummary::tbl_stack(list(tab_prop_1, tab_niv_1,tab_stat_1))
```

```r
tab_prop_2 <- tabl_fil(projet,24,"Anarchade","Propriétaire/Locataire",which(names(projet)=="q81"))
tab_niv_2 <- tabl_fil(projet,24,"Anarchade","Niveau d'instruction",which(names(projet)=="q25"))
tab_stat_2 <- tabl_fil(projet,24,"Anarchade","Statut juridique",which(names(projet)=="q12"))

tab_fil_2 <- gtsummary::tbl_stack(list(tab_prop_2, tab_niv_2,tab_stat_2))


tab_prop_3 <- tabl_fil(projet,25,"Arachide","Propriétaire/Locataire",which(names(projet)=="q81"))
tab_niv_3 <- tabl_fil(projet,25,"Arachide","Niveau d'instruction",which(names(projet)=="q25"))
tab_stat_3 <- tabl_fil(projet,25,"Arachide","Statut juridique",which(names(projet)=="q12"))

tab_fil_3 <- gtsummary::tbl_stack(list(tab_prop_3, tab_niv_3,tab_stat_3))

tab_prop_4 <- tabl_fil(projet,26,"Arachide","Propriétaire/Locataire",which(names(projet)=="q81"))
tab_niv_4 <- tabl_fil(projet,26,"Arachide","Niveau d'instruction",which(names(projet)=="q25"))
tab_stat_4 <- tabl_fil(projet,26,"Arachide","Statut juridique",which(names(projet)=="q12"))

tab_fil_4 <- gtsummary::tbl_stack(list(tab_prop_4, tab_niv_4,tab_stat_4))

tab_crois1 <- gtsummary::tbl_merge(
  list(tab_fil_1,tab_fil_2),
  tab_spanner = c("**Arachide**", "**Anacarde**") ## intitulé des groupes de tableau associés
)

tab_crois2 <- gtsummary::tbl_merge(list(tab_fil_3,tab_fil_4),tab_spanner = c("**Mangue**","**Riz**"))

tab_crois1%>% modify_caption("Repartition des PM produisant l'arachide et l'anacarde en fonction sexe,ni
```

Table 3: Repartition des PM produisant l'arachide et l'anacarde en fonction sexe,niveau d'instruction, Proprietaire et du statut juridique

	Arachide			**Anacarde**		
	Femme (n=93, 86%)	**Homme** (n=15, 14%)	**p-value**[1]	**Femme** (n=40, 66%)	**Homme** (n=21, 34%)	**p-value**[1]
Propriétaire/Locataire			0.4			0.2
Locataire	9 (8.3%)	3 (2.8%)		3 (4.9%)	4 (6.6%)	
Propriétaire	84 (78%)	12 (11%)		37 (61%)	17 (28%)	
Niveau d'instruction			0.3			<0.001
Aucun niveau	38 (35%)	5 (4.6%)		12 (20%)	1 (1.6%)	
Niveau primaire	20 (19%)	3 (2.8%)		15 (25%)	2 (3.3%)	
Niveau secondaire	30 (28%)	4 (3.7%)		9 (15%)	6 (9.8%)	
Niveau Superieur	5 (4.6%)	3 (2.8%)		4 (6.6%)	12 (20%)	
Statut juridique			0.016			0.002
Association	2 (1.9%)	0 (0%)		1 (1.6%)	2 (3.3%)	

[1] Fisher's exact test

Table 3: Repartition des PM produisant l'arachide et l'anacarde en fonction sexe,niveau d'instruction, Proprietaire et du statut juridique

	Arachide			Anacarde		
	Femme (n=93, 86%)	**Homme** (n=15, 14%)	**p-value**[1]	**Femme** (n=40, 66%)	**Homme** (n=21, 34%)	**p-value**[1]
GIE	70 (65%)	9 (8.3%)		27 (44%)	8 (13%)	
Informel	20 (19%)	3 (2.8%)		10 (16%)	2 (3.3%)	
SA	0 (0%)	2 (1.9%)		0 (0%)	2 (3.3%)	
SARL	0 (0%)	1 (0.9%)		1 (1.6%)	5 (8.2%)	
SUARL	1 (0.9%)	0 (0%)		1 (1.6%)	2 (3.3%)	

[1]Fisher's exact test

```
#kable(tab_crois1,caption = "Repartition des PM produisant l'arachide et l'anacarde en fonction sexe,ni

##%>% kable(format = "latex",caption = "Repartition des PM produisant l'arachide et l'anacarde en fonct

tab_crois2%>% modify_caption("Repartition des PM produisant la mangue et le riz en fonction sexe,niveau
```

Table 4: Repartition des PM produisant la mangue et le riz en fonction sexe,niveau d'instruction, Proprietaire et du statut juridique

	Mangue			Riz		
	Femme (n=68, 76%)	**Homme** (n=21, 24%)	**p-value**[1]	**Femme** (n=77, 84%)	**Homme** (n=15, 16%)	**p-value**[1]
Propriétaire/Locataire			0.7			>0.9
Locataire	8 (9.0%)	3 (3.4%)		8 (8.7%)	1 (1.1%)	
Propriétaire	60 (67%)	18 (20%)		69 (75%)	14 (15%)	
Niveau d'instruction			0.004			<0.001
Aucun niveau	22 (25%)	4 (4.5%)		10 (11%)	1 (1.1%)	
Niveau primaire	20 (22%)	4 (4.5%)		26 (28%)	0 (0%)	
Niveau secondaire	21 (24%)	4 (4.5%)		28 (30%)	4 (4.3%)	
Niveau Superieur	5 (5.6%)	9 (10%)		13 (14%)	10 (11%)	
Statut juridique			<0.001			<0.001
Association				0 (0%)	2 (2.2%)	
GIE	62 (70%)	11 (12%)		73 (79%)	4 (4.3%)	
Informel	3 (3.4%)	2 (2.2%)		1 (1.1%)	2 (2.2%)	
SA	1 (1.1%)	2 (2.2%)		0 (0%)	3 (3.3%)	
SARL	1 (1.1%)	5 (5.6%)		1 (1.1%)	4 (4.3%)	

[1]Fisher's exact test

Table 4: Repartition des PM produisant la mangue et le riz en fonction sexe,niveau d'instruction, Proprietaire et du statut juridique

	Mangue			Riz		
	Femme (n=68, 76%)	**Homme** (n=21, 24%)	**p-value**[1]	**Femme** (n=77, 84%)	**Homme** (n=15, 16%)	**p-value**[1]
SUARL	1 (1.1%)	1 (1.1%)		2 (2.2%)	0 (0%)	

[1]Fisher's exact test

```
#kable(tab_crois2,caption ="Repartition des PM produisant la mangue et le riz en fonction sexe,niveau d

##%>% kable(format = "latex",caption = "Repartition des PM produisant la mangue et le riz en fonction s
```

1.2.3. Les statistiques descriptives de notre choix sur les autres variables

Nous choisissons de faire les statistiques descriptives de la répartition des PME en fonction de la region.Il ressort des resultats que qu'il n'existe pas à Dakar et à Sedhiou de PME produisant de la mangue et de l'Arachide.Plus de la moitié des PME de l'arachide sont implenté à Diourbel et Thiès tandis les 50% des PME de l'anacarde sont de Fatikc et Ziguinchor.Pour la filière Mangue,les PME sont situés dans la region de Saint-Louis et Thiès.Pour ce qui est des PME evoluant dans la production du riz,les PME sont implentés à Thiès et à Ziguinchor.Toutes ces repartions inegalitaires s'expliquent fondamentalement par l'ineagel repartition pluviometrie favorisant l'accès facile à la matière première dans certains regions.

```
theme_gtsummary_compact(set_theme = TRUE, font_size = NULL)

## Format de la sortie
theme_gtsummary_printer(
  print_engine = "flextable",
  #c("gt", "kable", "kable_extra", "flextable", "huxtable", "tibble"),
  set_theme = TRUE
)
               # Créons le tableau 1 pour arachide

tbl_1 <- projet %>%select(region,filiere_1) %>%
  gtsummary::tbl_summary(
    include = c(region,filiere_1),
    by = filiere_1
  ,label = list(region ~ "Région"))%>%
  add_overall()%>%
  bold_labels() %>%
  italicize_levels()%>%modify_column_hide(c(stat_0,stat_1))

               # Créons le tableau 2 pour anacharde

tbl_2 <- projet %>%select(region,filiere_2) %>%
  gtsummary::tbl_summary(
    include = c(region,filiere_2),
    by = filiere_2
  ,label = list(region ~ "Région"))%>%
  add_overall()%>%
  bold_labels() %>%
  italicize_levels()%>%modify_column_hide(c(stat_0,stat_1))
```

```
              # Créons le tableau 3 pour mangue
tbl_3 <- projet %>%select(region,filiere_3) %>%
  gtsummary::tbl_summary(
    include = c(region,filiere_3),
    by = filiere_3
  ,label = list(region ~ "Région"))%>%
  add_overall()%>%
  bold_labels() %>%
  italicize_levels()%>%modify_column_hide(c(stat_0,stat_1))

              # Créons le tableau 4 pour riz

tbl_4 <- projet %>%select(region,filiere_4) %>%
  gtsummary::tbl_summary(
    include = c(region,filiere_4),
    by = filiere_4
  ,label = list(region ~ "Région"))%>%
  add_overall()%>%
  bold_labels() %>%
  italicize_levels()%>%modify_column_hide(c(stat_0,stat_1))

              #Mergeons les 4 tableaux ci-dessus en un seul tableau

gtsummary::tbl_merge(
  list(tbl_1,tbl_2,tbl_3,tbl_4),
  tab_spanner = c("Arachide", "Anacarde","Mangue","Riz")
  ## intitulé des groupes de tableau associés
)%>% modify_caption("Repartition des PM produisant la mangue,le riz,arachide et anacarde en fonction des
```

Table 5: Repartition des PM produisant la mangue,le riz,arachide et anacarde en fonction des régions

	Arachide	Anacarde	Mangue	Riz
Characteristic1, N = 108[1]**1**, N = 61[1]**1**, N = 89[1]**1**, N = 92[1]				
Région				
Dakar	0 (0%)	1 (1.6%)	0 (0%)	1 (1.1%)
Diourbel	33 (31%)	0 (0%)	1 (1.1%)	0 (0%)
Fatick	12 (11%)	21 (34%)	3 (3.4%)	4 (4.3%)
Kaffrine	8 (7.4%)	0 (0%)	5 (5.6%)	1 (1.1%)
Kaolack	20 (19%)	0 (0%)	7 (7.9%)	4 (4.3%)
Kolda	1 (0.9%)	5 (8.2%)	0 (0%)	4 (4.3%)
Saint-Louis	1 (0.9%)	0 (0%)	42 (47%)	0 (0%)
Sédhiou	0 (0%)	3 (4.9%)	0 (0%)	3 (3.3%)
Thiès	27 (25%)	0 (0%)	25 (28%)	32 (35%)
Ziguinchor	6 (5.6%)	31 (51%)	6 (6.7%)	43 (47%)

^{1}n (%)

1.3 Un peu de cartographie

Comme le pays concerné est le Sénégal et que nous voulons analyser la repartition spatiale , il est intereesaant de representer géographiquement le Sénégal et de proceder à son decoupage suivant la première repartition administrative qui est la region.Ainsi,nous utiliserons la fonction raster qui importera le sheapfiledu Sénégal depuis le site **GADM**.

```
Senegal <- raster::getData("GADM", country = "Senegal", level = 1)
centre<-coordinates(Senegal)
noms1<-Senegal$NAME_1
w.nb1 <- poly2nb(Senegal,row.names = noms1,queen=TRUE)
par(oma = c(0, 0, 0, 0), mar = c(0, 0, 1, 0))
plot(Senegal)
text(centre[,1],centre[,2],noms1,cex=.55)
#title("Carte du Sénégal avec les régions")
layoutLayer( sources = "Source:Calculs de l'auteur", frame = TRUE,
             col = "NA", scale = NULL)
```

Figure 1: Carte du Sénégal avec les régions

1.3.1. Transformer le data.frame en données géographiques

Pour transformer de type le data.frame projet en un l'objet nommé projet_mapen de classe sf, on utilise la fonction *st_as_sf()* de la manière suivante .

```
projet_map<- st_as_sf(projet, coords = c("gps_menlongitude", "gps_menlatitude"))
class(projet_map)##La fonction st_as_sf prend les paramètres la base et les cooordonées indiquuant la l
```

```
## [1] "sf"         "data.frame"
```

1.3.2. Faites une réprésentation spatiale des PME suivant le sexe

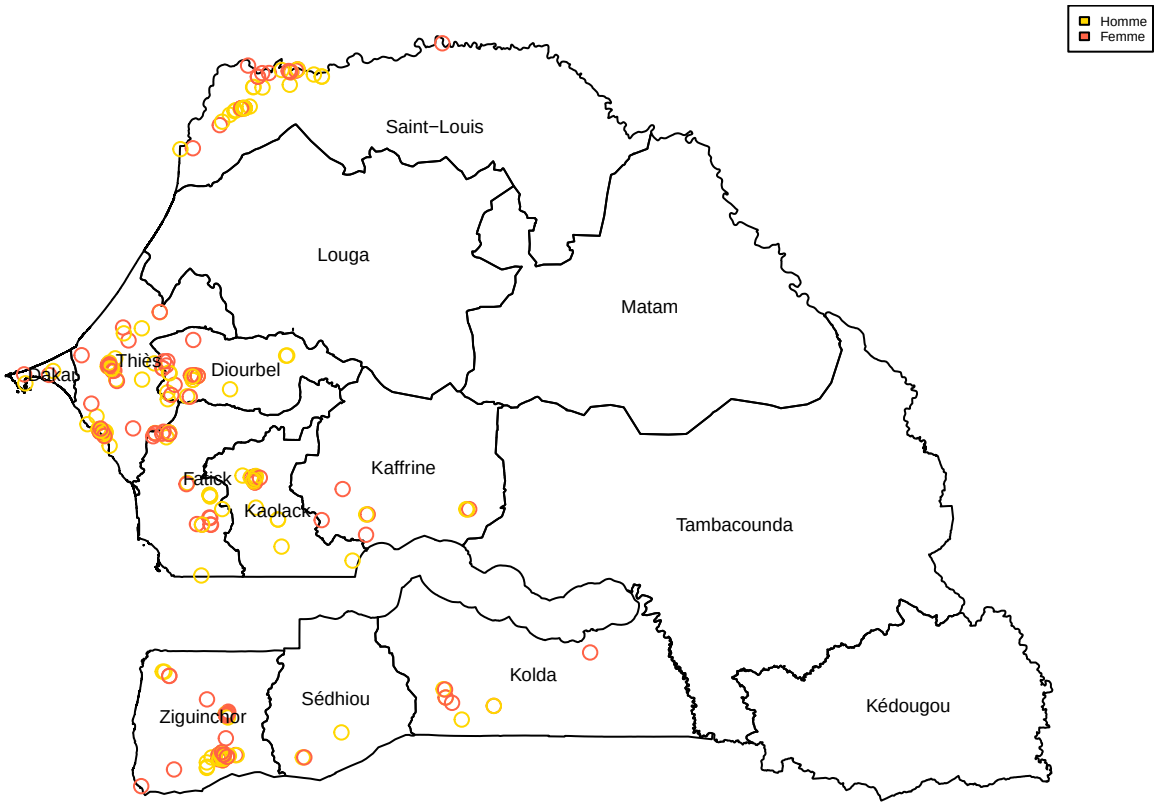
Tout comme les hommes proprietatires des PME repartis dans l'espace du pays,on retrouve de meme presque partout des PME des femmes repartit également dans l'espace sénégalais.Ce qui vient conforter l'idée la proposition égailité hommes-femmes

```
####GADM est un site ou est logé notre sheapefiles
Senegal <- raster::getData("GADM", country = "Senegal", level = 1)## level=1:recuperation des polygones
centre<-coordinates(Senegal)##Recupreation des coordonnées  du centre de chaque region
noms1<-Senegal$NAME_1##recuperation des noms des variables logéees dans var=NAME_1
w.nb1 <- poly2nb(Senegal,row.names = noms1,queen=TRUE)
par(oma = c(0, 0, 0, 0), mar = c(0, 0, 1, 0))
carte<-plot(Senegal)### generer la carte du,Sénégal
plot(projet_map["sexe"],col=c("gold","tomato"),add=TRUE)###plot la variable sexe de la base projet_map
attach(projet_map)
legend("topright",
       legend = c("Homme","Femme"),
       cex = 0.4,
       fill = c("gold","tomato"))###legende et les couleurs associées

text(centre[,1],centre[,2],noms1,cex=.55)##taille des noms de regions,de la carte et position au centre
#title("Repartition des PM en fonction du sexe")
layoutLayer( sources = "Source:Calculs de l'auteur", frame = TRUE,
             col = "NA", scale = NULL)
```

1.3.3. Faites une réprésentation spatiale des PME suivant le niveau d'instruction

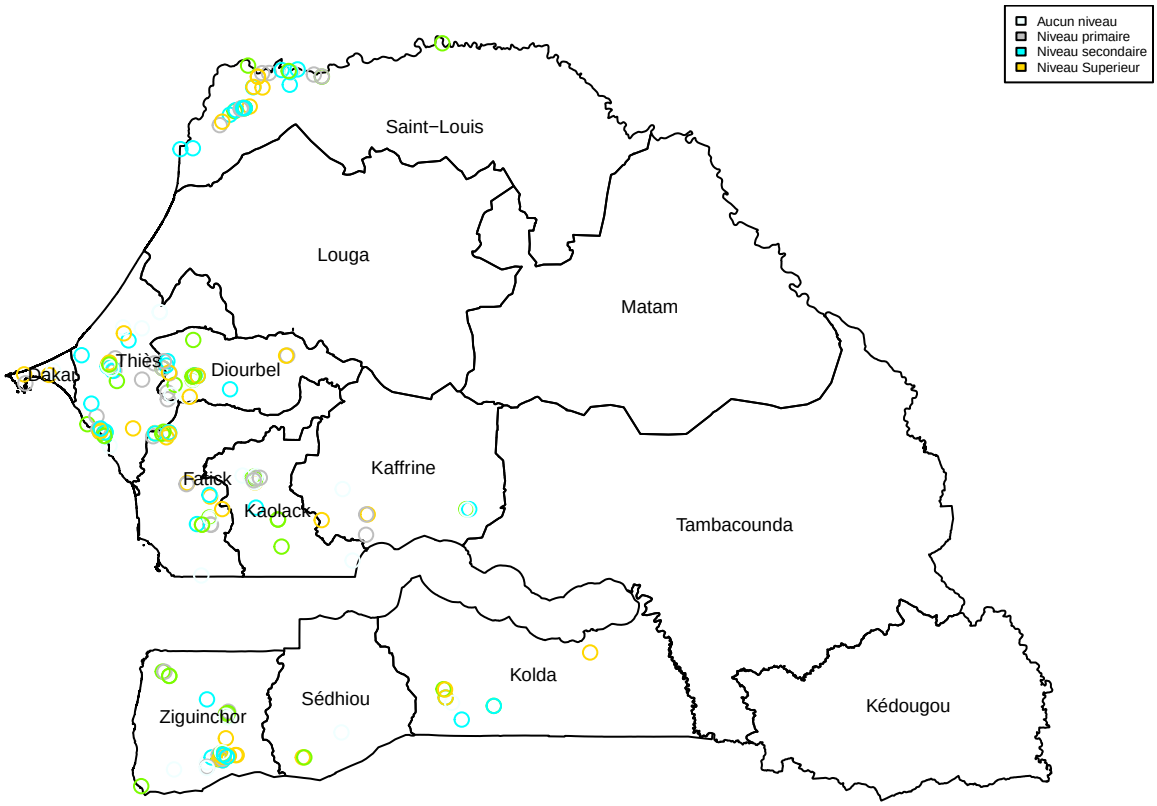
es regions du centre et Dakar presentent une heterogenité de la repartition des PM suivant le niveau d'instruction du CM.Quand aux regions de Kolda,les proprietaires des PME ont au moins le niveau Sécondaire.

```
Senegal <- raster::getData("GADM", country = "Senegal", level = 1)
centre<-coordinates(Senegal)
noms1<-Senegal$NAME_1
w.nb1 <- poly2nb(Senegal,row.names = noms1,queen=TRUE)
par(oma = c(0, 0, 0, 0), mar = c(0, 0, 1, 0))
carte<-plot(Senegal)
plot(projet_map["q12"],col=c("azure","gray","cyan","gold", "lawngreen"),add=TRUE)
attach(projet_map)
legend("topright",
       legend = c("Aucun niveau","Niveau primaire","Niveau secondaire","Niveau Superieur"),
       cex = 0.4,
       fill = c("azure","gray","cyan","gold",
                "lawngreen"))

text(centre[,1],centre[,2],noms1,cex=.55)
##title("Repartition spatiale des PM en fonction du niveau d'instruction du propiretaire ")
layoutLayer( sources = "Source:Calculs de l'auteur", frame = TRUE,
             col = "NA", scale = NULL)
```



Source:Calculs de l'auteur

Figure 2: Repartition spatiale des PM en fonction du sexe



Source:Calculs de l'auteur

Figure 3: Repartition spatiale des PM en fonction du niveau d'instruction du propiretaire

1.3.4. Faites une analyse spatiale de votre choix

Nous chosissons analyser spatialement la repartition des PM suivant le statut juridique.Ce choix s'explique par le fait que cette repartition nous donnera une idée du statut des PM suivant les differents regions du Sénégal.

Suivant l'espace, Dakar et Thiès representent la majorité des types de d'entreprise quel qu'en soit le statut juridique.Cela s'apercoit aisment si on sait que ces deux regions à elles seules regorenet les 80% des activités économiques du pays.Les regions du **centre(Kaolack,Fatick),Zinguichor et Saint Louis** presentent une repartition presque homogène des PME suivant le statut jurique.

```
Senegal <- raster::getData("GADM", country = "Senegal", level = 1)
centre<-coordinates(Senegal)
noms1<-Senegal$NAME_1
w.nb1 <- poly2nb(Senegal,row.names = noms1,queen=TRUE)
par(oma = c(0, 0, 0, 0), mar = c(0, 0, 1, 0))
carte<-plot(Senegal)
plot(projet_map["q12"],col=c("azure","gray","cyan","gold", "lawngreen","tomato","violet"),add=TRUE)
attach(projet_map)
legend("topright",
       legend = c("GIE","Informel","SUARL","SARL","Association","SA"),
       cex = 0.4,
       fill = c("azure","gray","cyan","gold",
                    "lawngreen","tomato","violet"))

text(centre[,1],centre[,2],noms1,cex=.55)
##title("Repartition spatiale des PM en fonction du niveau du statut juridique ")
layoutLayer( sources = "Source:Calculs de l'auteur", frame = TRUE,
             col = "NA", scale = NULL)
```

2.Partie

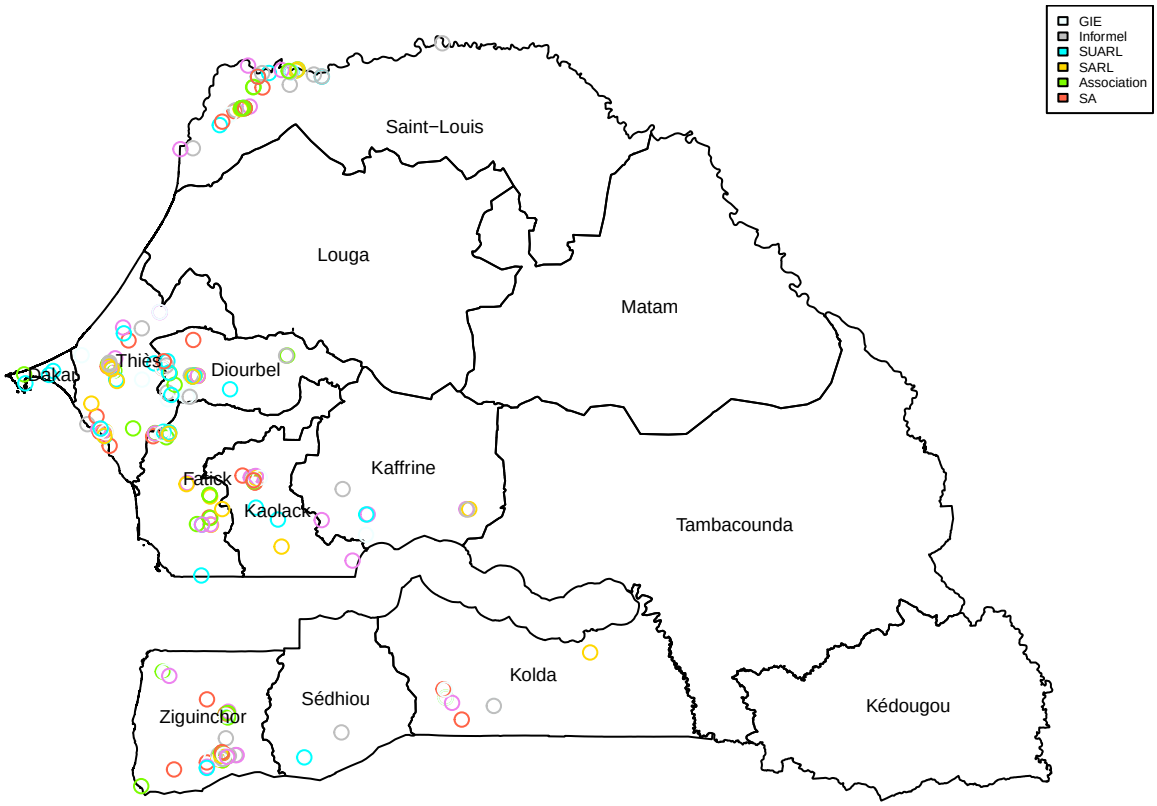
Dans cette partie, nous utiliserons principalement Le fichier excel **Base_Partie 2.xlsx** qui est une base de données artificielles.L'interet de cette partie resulte du fait qu'elle nous permet de passer en revue avec les methodes de renommer, creations des valeurs, impuations des variables,les tests statistques et la modelisations sur R.

2.1. Nettoyage et gestion des données

Nous importerons la base contenu dans un fechier de cette partie nommée **Base_Partie 2.xlsx** qui est un un ensemble de données artificielles .La specification du fichier est très important car s'il y'a beaucoup feuilles dans le classeur et que rien n'est specifier par defaut, il prendra la première feuille du classeur.

```
##Importation de la base de cette partie
partie2<- read_excel("Base_Partie 2.xlsx",sheet="data")
partie2<-data.frame(partie2)
head(partie2)%>%
  kable(format = "latex")
```

id	starttime	endtime	enumerator	district	age	sex	children_num	intention	country
2	2019-01-14 14:56:37	2019-01-14 15:11:10	6	1	33	1	1	1	
3	2019-01-14 16:12:22	2019-01-14 16:45:52	6	1	43	0	5	1	
4	2019-01-14 17:15:47	2019-01-14 17:45:47	6	1	28	0	0	1	
7	2019-01-14 13:04:51	2019-01-14 13:27:38	8	3	24	0	0	1	
8	2019-01-14 13:38:00	2019-01-14 14:31:16	8	3	29	0	0	1	
10	2019-01-14 15:52:17	2019-01-14 16:33:39	8	6	22	1	0	1	



Source:Calculs de l'auteur

Figure 4: Repartition spatiale des PM en fonction du niveau du statut juridique

2.1.1. Renommer la variable “country_destination” en “destination”

Il s’agit de renommer la variable en utilisant la fonction rename.Par suite, on imputera toutes les valeurs negatives par des NA.Une lecon tirée est que toutes les NA ne sont pas obligatoirement des valeurs manquantes.

```
##Renommons la variable "country_destination" en "destination"
partie2<-partie2 %>%
  rename(destination=country_destination)

## Deninissons les valeurs manquantes commevaleeurs manquantes
partie2$destination<-ifelse(partie2$destination<0,NA,partie2$destination)

##Verification s'il y'a des valeurs negatives
which(is.na(partie2$destination))
```

```
##  [1]  3 11 13 14 21 27 29 30 39 53 56 58 67 71 74 78 83 85 87 89
```

```
which(partie2$destination<0 )
```

```
## integer(0)
```

```
table(partie2$destination)
```

```
##
##  3  4  5  6  8  9 10 11 13
##  7  2  8  3 10 22 18  1  6
```

2.1.2. Créer une nouvelle variable contenant des tranches d’âge de 5 ans

Une simple tabulation nous revèle que la variable age a une modalité qui équivaut à 999.En s’inspirant de la socité et des conniassances demographiques, cette modalité est une anomalie d’autant plus la personne la plus agée aurait 128 selon la **Nouveelle Tribune**.Ainsi on procède à l’imputation de cette valeurs et à toutes les valeurs manquantes de la base.

```
attach(partie2)
##Recuperation du premier quartile
a<-quantile(partie2$age)[2]
##Recuperation du premier quartile
b<-quantile(partie2$age)[4]
##calcul de la borne inferieur
born_inf=a-1.5*(b-a)
##calcul de la borne superieur
born_sup=b+1.5*(b-a)

## Detection et imputation des valeurs aberrantes par la moyenne des ages
partie2$age_aberr<-ifelse((partie2$age<born_inf)|(partie2$age>born_sup),mean(partie2$age),partie2$age)
```

Nous allons à present recoder la variable age en une variable categorielle.Commme elle est la variable age propre.Dans la suite,nous l’utuliserons pour les besoins d’analyse.

```
attach(partie2)
##creation des bornes de la tranche d'age
ecart<-5
bornes<-seq(min(partie2$age_aberr),max(partie2$age_aberr),by=ecart)

##decoupage de la variable age des tranches d'age
partie2$age_categ<-cut(partie2$age_aberr,breaks = bornes)
```

Table 6: Effectifs des repondnats par tranches d'age

Var1	Freq
(15,20]	20
(20,25]	34
(25,30]	22
(30,35]	10
(35,40]	10

```
table(partie2$age_categ)%>% kable(format = "latex",caption ="Effectifs des repondnats par tranches d'ag
```

2.1.3. Créer une nouvelle variable contenant le nombre d'entretiens réalisés par chaque agent recenseur.

Cette question recommande qu'on regroupe d'abord les enqueteurs par groupe homogènes et après et recupere le nombre total d'entretien par enqueteur qui sera affecté à sa ligne dès qu'il porte le meme numero.

```
attach(partie2)
##La fonction group_by regroupe les enqueteurs par groupe homogène
partie2<-partie2%>% group_by(enumerator)%>%
  mutate(nbre_entretien=n())%>% distinct()## creation du nombre d'entretiens en distinctinguant les enq
```

2.1.4. Créer une nouvelle variable qui affecte aléatoirement chaque

Il s'agit de generer une variable aleatoire qui affectera à chaque personne une valeur 0 ou 1.Pour cela, nous utilisons la fonction sample.

```
set.seed(123)##fixation du générateur de nombres aléatoires
partie2<-partie2%>%rowwise()%>%##affectation de la base partie à elle meme en enrecuperant son nombre d
  mutate(regroup=sample(c(0,1),1))## creation de la variable aléatoire logée dans la base partie2
```

2.1.5. Fusionner la taille de la population de chaque district

Lobjectif est que que toutes les personnes interrogées aient une valeur correspondante représentant la taille de la population du district dans lequel elles vivent.On utilisera la fonction merge avec comme clé d'identification la variable district.

```
## Importation de la feuille
partie2_f2<- read_excel("Base_Partie 2.xlsx",sheet="district")
partie2_f2<-data.frame(partie2_f2)
## fusion des deux bases
partie2_taill<-partie2%>%
               merge(partie2_f2,by="district")## la fonction merge dans notre cas fusionne la base langt
```

2.1.6. Calculer la durée de l'entretien et indiquer la durée moyenne de l'entretien par enquêteur

On utilise la library lubridate dans laquelle est logée la fonction permettant de calculer la durée entre deux date.Par la suite, nous procederons au caclu de la durée moyenne ˆpar enqueteeur

```
library(lubridate)##library de calcul des dates
 partie2_taill<-partie2_taill %>%
   mutate(
     ###la fonction permet de calculer la duree comprise entre deux dates
     dure_entr = time_length(
```

Table 7: Durée moyenne en minutes d'interview par enqueteur

enumerator	moyenne
6	25.84667
14	25.56111
11	33.48333
20	28.76852
18	36.85833
13	31.59583
4	36.48333
1	68.14667
12	48.16667
8	40.13056
15	28.65000
9	114.76667
10	55.27667
5	33.55833
17	29.28611
7	37.16429

```
      interval(
         start = starttime,
         end = endtime
         ),
       unit = "hour"
       )
     )
   #select(nom, date_naissance, age) %>%
   #glimpse()

 colnames(partie2_taill)
```

```
##  [1] "district"       "id"             "starttime"      "endtime"
##  [5] "enumerator"     "age"            "sex"            "children_num"
##  [9] "intention"      "destination"    "age_aberr"      "age_categ"
## [13] "nbre_entretien" "regroup"        "population"     "dure_entr"
```

2.1.7. calcul de la durée moyenne en minutes par enqueteur

```
partie2_taill%>% group_by(enumerator)%>%
  transmute(moyenne=mean(dure_entr)*60)%>% distinct()%>%
    kable(format = "latex",caption ="Durée moyenne en minutes d'interview par enqueteur")
```

2.1.8. Renommez toutes les variables de l'ensemble de données en ajoutant le préfixe"endline_"

On renomme les variables en ajoutant le prefixes.Pour cela, on utilise une boucle avec la fonction for qui parcourt les colonnes de la base et ajoute ce prefixe.

```
## recueillir le nombre de colonne de la base
n<-ncol(partie2_taill)

##une boucle qui va parcourir le nombre de ligne pour ajouter les prefixes
```

```
for (i in 1:n) {
  colnames(partie2_taill)[i]<-paste("endline_",colnames(partie2_taill)[i],sep ="" )

}
```

2.2. Analyse et visualisation des données

2.2.1.Créez un tableau récapitulatif contenant l'âge moyen et le nombre moyen d'enfants par district.

Il ressort que le district 1 presente l'age moyen des repondnats le plus élevé avec un age moyen de 28.7 tandis que le district 6 presente l'age moyen le plus faible avec 23.3.Autre remarque, les ages moyens inter-districts sont très peu dispersées,chose qui revèle et tourne de 24.Cela est une caracteristique phare de la popualtion de l'Afrique de l'ouest fortement jeune.

En ce qui concerne le nombre d'enfants, on note des nombres moyen d'enfants relativement très faibles avec un nombre moyen de 1.5 enfants et toujours relevé dans le district1.De meme, il ressort que dans les districts 3 et 4,le nombre moyen des repondants est de 0.

```
## tableaude l'age moyen en fonction du district
calcul_ag<-partie2_taill %>%
  gtsummary::tbl_continuous(
    variable = endline_age_aberr, ## variable à représenter

    statistic = ~"{mean}", ## stat à  calculer

    include =endline_district , ## variables pour lesquelles ou suivant lesquelles on calcul
    digits = ~1 ## formatage des chiffres
  ) %>%
  modify_header(
      list(
        all_stat_cols() ~ "**age du repondant** \n(n={n}, {style_percent(p)}%)"
      ))

###

calcul_enfant<-partie2_taill %>%
  gtsummary::tbl_continuous(
    variable = endline_children_num, ## variable à représenter

    statistic = ~"{mean}", ## stat à  calculer

    include =endline_district , ## variables pour lesquelles ou suivant lesquelles on calcul
    digits = ~1 ## formatage des chiffres
  ) %>%
  modify_header(
      list(
        all_stat_cols() ~ "**nombre d'enfant du repondant** \n(n={n}, {style_percent(p)}%)"
      ))
tbl_merge(list(calcul_ag,calcul_enfant),tab_spanner = c("" ,""))%>% modify_caption("tableau récapitulati
```

Table 8: tableau récapitulatif contenant l'âge moyen et le nombre moyen d'enfants par district

Characteristicage du repondant (n=97, 100%)[1]	**nombre d'enfant du repondant** (n=97, 100%)[2]	
endline_district		
1	28.7	1.5
2	26.9	0.9
3	26.1	0.0
4	26.0	0.0
5	24.3	0.5
6	23.2	0.1
7	26.9	0.2
8	24.6	1.3

[1] endline_age_aberr: Mean

[2] endline_children_num: Mean

2.2.2.Testez si la différence d'âge entre les sexes

La differnece entre l'age moyen des hommes et et celui des hommes est de celle des femmes est de 2.6% et cette difference est significative à 20%.On conclut donc que ladifference de l'age moyenne entre homme et celui des hommes n'est pas significatif. est statistiquement significative au niveau de 5 %.

```
partie2_taill %>%
  gtsummary::tbl_summary(

    include = endline_age_aberr, ##variables à representer dans le tableau

    statistic =endline_age_aberr  ~ "{mean} ", ## stat associée à ces variables

    by = endline_sex ## groupes
  )%>%  add_difference()%>% modify_caption("Test si la différence d'âge entre les sexes")
```

Table 9: Test si la différence d'âge entre les sexes

Characteristic	**0**, N = 86[1]	**1**, N = 11[1]	**Difference**[2]	**95% CI**[2,3]	**p-value**[2]
endline_age_aberr	25.8	23.2	2.6	-1.9, 7.1	0.2

[1] Mean

[2] Welch Two Sample t-test

[3] CI = Confidence Interval

2.2.3. Créer un nuage de points de l'âge en fonction du nombre d'enfants

Le graphique ci-dessous montre que que l'age du repondant n'est une fonction lineaire du nombre d'enfant qu'il a .On note une grande dispersion entre l'age du repondant et le nombre d'enfants.Ceci peut s'expliquer du fait qu'il y'a des jeunes qui ne pratiquent pas la planning familia et par ricochet se voit avec un nombre élevé des enfants à la fleur de l'age tandis certains ont eu appliqués le planing familial à leur jeunesse et se triuve avecpeu d'enfant meme aagé et vice-versa.Tout cela laisse transparaitre la problematique d'acceptation du planning par tous familial en Afrique de l'Ouest.

```
library(ggplot2)
library()
nuage<-ggplot(partie2_taill) +
    geom_point(aes(x = endline_age_aberr, y = endline_children_num))

 # ggMarginal(nuage,type="densigram")

    #ggMarginal(nuage,type="boxplot")
```

2.2.4.La variable "intention" indique si les migrants potentiels ont l'intention de migrer sur une échelle de 1 à 7.

Estimez l'effet de l'appartenance au groupe de traitement sur l'intention de migrer.

Nous avons utilisé le modèle logistique multinomial pour estimer l'effet de d'appartenenance àl'intention de migrer ou pas.Ce choix du modèle s'explique par deux raisons principales:

La premièreraison est liée au resultat attendu .Il s'agit de calculer la chance des personnes tirées aléatoirement d'appartenir une categorie d'intention de migrer. Donc le modèle logistiuqe avec les ordres rations devient incontournable.

La deuxième raison est la nature de la variable à expliquer qui contient 6 modalités et donc le modèle logistiquebinomiale apparait plus adapté à ces types d'estimations meilleur.

Les resultats du modèle restent non predictif pour tous les élements car la p-value superieur à 5%.Ainsi, les resultats du modèle ne sont pas significatifs pour toutes les modalités de la variable intentiontion.Ainsi donc le modèle n'est pas predictif pour ses echelles.Par consequent nous nous reservons de commenter les ordres rations données.Tout ce qu'on peut noter,c'est qu'une personne prise aléatoirement, on ne peut pas predire son intention de migrer suivant les échelles definies.

```
## Utilisation du modèle logistique

library(nnet)
regression<-multinom( endline_intention ~ endline_regroup, data = partie2_taill)
```

```
## # weights:  21 (12 variable)
## initial  value 188.753284
## iter  10 value 116.109117
## iter  20 value 115.901772
## final  value 115.901310
## converged
```

```
tbl_regression(regression,exponentiate=TRUE)%>% modify_caption("Estimation l'effet de l'appartenance au
```

Table 10: Estimation l'effet de l'appartenance au groupe de traitement sur l'intention de migrer

Outcome	Characteristic	OR[1]	95% CI[1]	p-value
2	endline_regroup	0.47	0.05, 4.82	0.5
3	endline_regroup	0.61	0.14, 2.58	0.5
4	endline_regroup	3.56	0.64, 19.8	0.15
5	endline_regroup	1.42	0.27, 7.61	0.7
6	endline_regroup	5.69	0.60, 53.9	0.13
7	endline_regroup	0.00	0.00, Inf	>0.9

[1]OR = Odds Ratio, CI = Confidence Interval

2.2.5.Créez un tableau de régression avec 3 modèles.

La variable de résultat est toujours "intention". Modèle A : Modèle vide - Effet du traitement sur les intentions. Modèle B : Effet du traitement sur les intentions en tenant compte de l'âge et du sexe. Modèle C : Identique au modèle B mais en contrôlant le district. Les résultats des trois modèles doivent être affichés dans un seul tableau.

Comme les resultats avancées dans la question predente, nous utiliserons le modèle logistique pour les 3 modèles. Les resultats du modèle A sont les memes que celui de la question precedente. Concerant le modèle B,les resultats sont significatifs à 5% pour l'echelle de 3 et 5 de l'intention de migrer.En fait, en fonction de son age, un repondant presente 1.06 fois plus de chances d'avoir l'intention de migrer avec comme echelle 1 ou 2.Quant en fonction du sexe, le modèle reste significatif à 5% pour les echelles 4,5,6,7 de l'intenetion de migrer mais ne presente aucune par rapport sa classe . Le modèle B, en analysant une personne prise aléatoirement et en fonction de son district, le modèle est seulement predictif à l'echelle 3 et 5 de l'intention de migrer au seuil de 5%.En effet, il ressort que ces personnes presentent 1.16 et 1.30 fois plus de chances d'avoir l'intention de migrer respectivement pour les echelles 3 et 5. En fonction

```
##Modèle A :nous ferons une regression mutinomiale de intention en fonction de la taille du meange

modeleA<-multinom( endline_intention ~ endline_regroup, data = partie2_taill)
```

```
## # weights:  21 (12 variable)
## initial  value 188.753284
## iter  10 value 116.109117
## iter  20 value 115.901772
## final  value 115.901310
## converged
```

```
regA<-tbl_regression(modeleA,exponentiate=TRUE)

##modèle B
modeleB<-multinom( endline_intention ~ endline_age_aberr+endline_sex, data = partie2_taill)
```

```
## # weights:  28 (18 variable)
## initial  value 188.753284
## iter  10 value 116.859566
```

```
## iter  20 value 116.307380
## iter  30 value 116.263638
## final  value 116.262690
## converged
```

```
regB<-tbl_regression(modeleB,exponentiate=TRUE)

##modèle C
modeleC<-multinom( endline_intention ~ endline_regroup+endline_district, data = partie2_taill)
```

```
## # weights:  28 (18 variable)
## initial  value 188.753284
## iter  10 value 115.175644
## iter  20 value 114.052987
## iter  30 value 114.047770
## final  value 114.047749
## converged
```

```
regC<-tbl_regression(modeleC,exponentiate=TRUE)

tbl_merge(list(regA,regB,regC))%>% modify_caption("tableau de régression avec 3 modèles")
```

Table 11: tableau de régression avec 3 modèles

	Table 1				Table 2				Table 3			
Characteristic	**Outcome**	**OR[1]**	**95% CI[1]**	**p-value**	**Outcome**	**OR[1]**	**95% CI[1]**	**p-value**	**Outcome**	**OR[1]**	**95% CI[1]**	**p-value**
endline_regroup	2	0.47	0.05, 4.82	0.5					2	0.54	0.05, 5.63	0.6
endline_regroup	2	0.47	0.05, 4.82	0.5					3	0.56	0.13, 2.41	0.4
endline_regroup	2	0.47	0.05, 4.82	0.5					4	3.40	0.60, 19.1	0.2
endline_regroup	2	0.47	0.05, 4.82	0.5					5	1.26	0.23, 6.88	0.8
endline_regroup	2	0.47	0.05, 4.82	0.5					6	5.70	0.59, 54.9	0.13
endline_regroup	2	0.47	0.05, 4.82	0.5					7	0.00	0.00, 0.00	<0.001
endline_regroup	3	0.61	0.14, 2.58	0.5					2	0.54	0.05, 5.63	0.6
endline_regroup	3	0.61	0.14, 2.58	0.5					3	0.56	0.13, 2.41	0.4
endline_regroup	3	0.61	0.14, 2.58	0.5					4	3.40	0.60, 19.1	0.2
endline_regroup	3	0.61	0.14, 2.58	0.5					5	1.26	0.23, 6.88	0.8
endline_regroup	3	0.61	0.14, 2.58	0.5					6	5.70	0.59, 54.9	0.13
endline_regroup	3	0.61	0.14, 2.58	0.5					7	0.00	0.00, 0.00	<0.001
endline_regroup	4	3.56	0.64, 19.8	0.15					2	0.54	0.05, 5.63	0.6
endline_regroup	4	3.56	0.64, 19.8	0.15					3	0.56	0.13, 2.41	0.4
endline_regroup	4	3.56	0.64, 19.8	0.15					4	3.40	0.60, 19.1	0.2
endline_regroup	4	3.56	0.64, 19.8	0.15					5	1.26	0.23, 6.88	0.8
endline_regroup	4	3.56	0.64, 19.8	0.15					6	5.70	0.59, 54.9	0.13
endline_regroup	4	3.56	0.64, 19.8	0.15					7	0.00	0.00, 0.00	<0.001

[1]OR = Odds Ratio, CI = Confidence Interval

Table 11: tableau de régression avec 3 modèles

	Table 1				Table 2				Table 3			
Characteristic	**Outcome**	**OR**[1]	**95% CI**[1]	**p-value**	**Outcome**	**OR**[1]	**95% CI**[1]	**p-value**	**Outcome**	**OR**[1]	**95% CI**[1]	**p-value**
endline_regroup	5	1.42	0.27, 7.61	0.7					2	0.54	0.05, 5.63	0.6
endline_regroup	5	1.42	0.27, 7.61	0.7					3	0.56	0.13, 2.41	0.4
endline_regroup	5	1.42	0.27, 7.61	0.7					4	3.40	0.60, 19.1	0.2
endline_regroup	5	1.42	0.27, 7.61	0.7					5	1.26	0.23, 6.88	0.8
endline_regroup	5	1.42	0.27, 7.61	0.7					6	5.70	0.59, 54.9	0.13
endline_regroup	5	1.42	0.27, 7.61	0.7					7	0.00	0.00, 0.00	<0.001
endline_regroup	6	5.69	0.60, 53.9	0.13					2	0.54	0.05, 5.63	0.6
endline_regroup	6	5.69	0.60, 53.9	0.13					3	0.56	0.13, 2.41	0.4
endline_regroup	6	5.69	0.60, 53.9	0.13					4	3.40	0.60, 19.1	0.2
endline_regroup	6	5.69	0.60, 53.9	0.13					5	1.26	0.23, 6.88	0.8
endline_regroup	6	5.69	0.60, 53.9	0.13					6	5.70	0.59, 54.9	0.13
endline_regroup	6	5.69	0.60, 53.9	0.13					7	0.00	0.00, 0.00	<0.001
endline_regroup	7	0.00	0.00, Inf	>0.9					2	0.54	0.05, 5.63	0.6
endline_regroup	7	0.00	0.00, Inf	>0.9					3	0.56	0.13, 2.41	0.4
endline_regroup	7	0.00	0.00, Inf	>0.9					4	3.40	0.60, 19.1	0.2
endline_regroup	7	0.00	0.00, Inf	>0.9					5	1.26	0.23, 6.88	0.8
endline_regroup	7	0.00	0.00, Inf	>0.9					6	5.70	0.59, 54.9	0.13
endline_regroup	7	0.00	0.00, Inf	>0.9					7	0.00	0.00, 0.00	<0.001
endline_age_aberr					2	1.00	0.85, 1.19	>0.9				
endline_age_aberr					3	1.06	0.95, 1.18	0.3				
endline_age_aberr					4	1.06	0.94, 1.20	0.4				
endline_age_aberr					5	1.00	0.86, 1.15	>0.9				
endline_age_aberr					6	0.99	0.84, 1.16	0.9				
endline_age_aberr					7	0.94	0.71, 1.23	0.6				
endline_sex					2	2.02	0.18, 22.4	0.6				
endline_sex					3	0.77	0.08, 6.99	0.8				
endline_sex					4	0.00	0.00, 0.00	<0.001				
endline_sex					5	0.00	0.00, 0.00	<0.001				
endline_sex					6	0.00	0.00, 0.00	<0.001				
endline_sex					7	0.00	0.00, 0.00	<0.001				
endline_district									2	0.85	0.52, 1.38	0.5
endline_district									3	1.16	0.87, 1.56	0.3

[1]OR = Odds Ratio, CI = Confidence Interval

Table 11: tableau de régression avec 3 modèles

	Table 1			Table 2			Table 3		
Characteristic	**Outcome**	**OR**[1]	**95% CI**[1] **p-value**	**Outcome**	**OR**[1]	**95% CI**[1] **p-value**	**Outcome**	**OR**[1]	**95% CI**[1] **p-value**
endline_district							4	1.08	0.75, 1.54 0.7
endline_district							5	1.30	0.87, 1.93 0.2
endline_district							6	1.00	0.65, 1.53 >0.9
endline_district							7	1.21	0.66, 2.22 0.5

[1]OR = Odds Ratio, CI = Confidence Interval

Partie 3

Dans cette partie nous utiliserons la base **ACLED-Western_Africa.csv**. C'est une base sur les evenements (violences politiques, marches, ...) en Afrique de l'Ouest.

3.1. Faisons une application r shiny permettant

3.1.1. visualisation des évenements par pays (le nombre d'evenement par pays dans une carte)

3.1.2. visualisation des évenements par pays, type, annee et la localisation

Conclusion

Au sorti de ces travaux pratiques, nous notons que le logiciel R nous offre un évantail de possibilité pour traiter et analyser les données.De meme, R semble etre efficace dans les analyses spatiales des informations géographiques.Nous terminerons par l'opportunité que nous offre ce logiciel dans la modelisation et la redaction des rapports sous format pdf et word à generer automatiquement sans pour autant perdre du temps à la mise en forme des tableaux et tableaux.Ce qui reconforte l'idée et les objectifs de ce module dans le programme.

BIBLIOGRAPHIE

- http://www.cef-cfr.ca}
- http://www.math-evry.cnrs.fr
- INTRODUCTION À L'ENVIRONNEMENT DE PROGRAMMATION STATISTIQUE R Y. BROSTAUX
- Philippe Grosjean, Frédéric Ibanez : Manuel de l'utilisateur de la librairie de Fonctions pour R et pour S+
- Faouzi LYAZRHI , Une Introduction au langage R