Contents

REPUBLIQUE DU SENEGAL

Un Peuple-Un But-Une Foi

MINISTERE DE L'ECONOMIE DU PLAN ET DE LA COOPERATION

AGENCE NATIONALE DE LA STATISTIQUE ET DE LA DEMOGRAPHIE

****

ECOLE NATIONALE DE LA STATISTIQUE ET DE L'ANALYSE ECONOMIQUE

ENSAE-PIERRE NDIAYE

PROJET DE STATISTIQUE SUR R

THEME :

PROJET VISANT A METTRE EN APPLICATION TOUT CE QUI A ETE VU PENDANT LE DE COURS DE R EN ISE1 ANNEE 2022/2023

Rédigé par :

TANGOUO KUETE Ivana

Elèves ingénieur statisticien économiste à

L'ENSAE de Dakar

Sous la supervision de :

M. HEMA Haboubacar

Ingénieur de Travaux statisque

Annee academique 2022-2023

Table des matières

PARTIE 1

bibliothèques necessaires

```
library(readxl)   ## Importer des fichiers avec extenxion xlsx
library(gtsummary) ## importer gtsummary
library("dplyr")
library(flextable)
library(leaflet) ## creation de la carte
library(sf)       ## pour la cartographie
library(rnaturalearth)  ## Importer les données cartographique
library(sp)
library(kableExtra)         ## Faire sortir les tableau sous forme de tableau
library(knitr)
library(htmlwidgets)        ## Creer des fichiers sous extension htlm
library(webshot)            ## Faire la capture d'écran suivant le format d'mage souhaité
library("lubridate")       ## gestion des dates
library(ggplot2)           ## tracé des graphes
library(ggExtra)           #
library(nnet)
library(GGally)
library(effects)
library(gridExtra)
library(forcats)
```

1- Préparation des données

1.1 Description

1.2 Importation et mise en forme

1.2.1 Importer la base de données dans un objet de type data.frame nommé projet

On importe la librairies readlx pour importer la base excel

```
projet <- read_excel("Base_Partie 1.xlsx",
                     range = NULL,
                    col_names = TRUE,
                     col_types = NULL)
```

1.2.2 Selection les variables mentionnees dans la section description.

Ici on a plus besoin de sélectionner les variables puisqu'elle sont déja sélectionner lors de l'importation de la base.

1.2.3 Faites un tableau qui resume les valeurs manquantes par variable

Il est important d'importer la librairie dplyr

```
val_manque<-data.frame(nbre_valeur_manquantes=colSums(is.na(projet)),
                       proportion=colSums(is.na(projet))*100/nrow(projet))%>%
                      dplyr::mutate(proportion=paste0(proportion, "%"))

kable(val_manque, format="latex")
```

	nbre_valeur_manquantes	proportion
key	0	0%
q1	0	0%
q2	0	0%
q23	0	0%
q24	0	0%
q24a_1	0	0%
q24a_2	0	0%
q24a_3	0	0%
q24a_4	0	0%
q24a_5	0	0%
q24a_6	0	0%
q24a_7	0	0%
q24a_9	0	0%
q24a_10	0	0%
q25	0	0%
q26	0	0%
q12	0	0%
q14b	1	0.4%
q16	1	0.4%
q17	131	52.4%
q19	120	48%
q20	0	0%
filiere_1	0	0%
filiere_2	0	0%
filiere_3	0	0%
filiere_4	0	0%
q8	0	0%
q81	0	0%
gps_menlatitude	0	0%
gps_menlongitude	0	0%
submissiondate	0	0%
start	0	0%
today	0	0%

1.2.4 Valeurs manquantes pour la variable key

```
PME_manquant<-projet %>% filter(key=="NA")
kable(PME_manquant[,1:6])
```

key	q1	q2	q23	q24	q24a_1

1.3 Création de variables

1.3.1 Renommer les variables

Renommer q1 en region,q2 en departement,q23 en sexe

```
projet <-projet %>% dplyr::rename(region=q1,departement=q2,sexe=q23)
```

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

```
projet$sexe_2 <- ifelse(projet$sexe == "Femme", 1,0 )

## Placer la nouvelle variable créer près de la variable sexe
projet <- projet%>% relocate(sexe_2, .after = sexe)

# affichage
kable(projet[1:3,1:6],format="latex")
```

key	region	departement	sexe	sexe_2	q24
uuid:68bff42b-1228-4c66-9bcc-e6d312d9fea6	Diourbel	Bambey	Femme	1	65
uuid:d70b3c7e-3ca0-4358-bc59-3f7f6baf55e9	Thiès	Mbour	Femme	1	52
uuid:0ac18b64-7d85-4bb9-a842-698ac79909af	Thiès	Mbour	Femme	1	65

1.3.3 Créer un data.frame nommé langues qui prend les variables key et les variables correspondantes décrites plus haut.

```
langues<-projet %>% dplyr::select(key,starts_with("q24a_"))
kable(langues[1:5,1:6],format="latex")
```

key	q24a_1	q24a_2	q24a_3	q24a_4	q24a_5
uuid:68bff42b-1228-4c66-9bcc-e6d312d9fea6	0	1	0	1	0
uuid:d70b3c7e-3ca0-4358-bc59-3f7f6baf55e9	1	1	0	0	1
uuid:0ac18b64-7d85-4bb9-a842-698ac79909af	1	1	0	0	0
uuid:c52cf5e4-8c28-4e65-998b-3fe2a971a1a3	1	1	0	0	1
uuid:ac177870-001c-4ada-8747-c22ffe4e4596	1	1	1	0	0

1.3.4 Créer une variable parle qui est égale au nombre de langue parlée par le dirigeant de la PME.

```
langues$parle<-rowSums(langues[, -1])
unlist(colnames(langues))
```

```
##  [1] "key"     "q24a_1"  "q24a_2"  "q24a_3"  "q24a_4"  "q24a_5"  "q24a_6"
##  [8] "q24a_7"  "q24a_9"  "q24a_10" "parle"
```

```
kable(langues[1:5,2:11],format="latex")
```

q24a_1	q24a_2	q24a_3	q24a_4	q24a_5	q24a_6	q24a_7	q24a_9	q24a_10	parle
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	2
1	1	0	0	1	0	0	0	0	3
1	1	1	0	0	1	0	0	0	4

1.3.5 Merger les data.frame projet et langues

```
Final<-merge(projet, langues, by = "key")
```

2 Analyses descriptives

2.1 Repartition des PME

repartition par sexe, niveau d'instruction et statut juridique: Nous allons importer la bibliothèque gtsummary.

dans notre base nous avons 250 PME , dont 76% des dirigeants sont des femmes contre 24% dirigeants hommes.

s

```
f<-Final %>%
   tbl_summary(
                  label = c(q25 ~ "Niveau d'inst",
                  q12~"Statut juridique",
                  q81~"Propriétaire "),
                  include=c("sexe","q25","q12","q81"),
                  missing="no" ) %>%
   modify_header(
    list(
      label ~ "Variable",
      all_stat_cols(stat_0 = FALSE) ~ "_{level}_ (n={n}, {style_percent(p)}%)",
      stat_0 ~ "TOTAL ={N}"
    )
  )
```

2.1 Grand tableau de fusion

Ici nous allons tout d'abord creer 04 tableaux avec gtsummary qui nous donne la repartion suivant les filières croisé avec le genre . Nous allons par la suite merger ces 04 tableaux avec la fonction tbl_merge.

tableau de la filière d'arachide

```
t1<-Final %>%
  dplyr::filter(filiere_1=="1") %>%
  mutate(filiere_1 = recode(filiere_1, `1` = "fil d'arachide")) %>%
  dplyr::select(sexe, q25, q12, q81) %>%
        gtsummary::tbl_summary(
      label = list(q25 ~ "Niveau d'inst",q12 ~ "Statut juridique",
            q81 ~ "Propriétaire ",sexe~"sexe"),
          missing = "no"
        ) %>% modify_header(label = "**Variable**") %>%
  bold_labels()
```

Tableau filiere de l'anarcarde

```
t2<-Final %>%
  dplyr::filter(filiere_2=="1") %>%
  mutate(filiere_2 = recode(filiere_2, `1` = "fil d'anacarde")) %>%
  dplyr::select(sexe, q25, q12, q81) %>%
  gtsummary::tbl_summary(
    label = list(q25 ~ "Niveau d'inst",q12 ~ "Statut juridique",
            q81 ~ "Propriétaire ",sexe~"sexe"),
          missing = "no"
        ) %>% modify_header(label = "**Variable**") %>%
  bold_labels()
```

Tableau filiere mangue

```
t3<-Final %>%
  dplyr::filter(filiere_3=="1") %>%
  mutate(filiere_3= recode(filiere_3, `1` = "fil de mangue")) %>%
  dplyr::select(sexe, q25, q12, q81) %>%
  gtsummary::tbl_summary(
      label = list(q25 ~ "Niveau d'inst",q12 ~ "Statut juridique",
            q81 ~ "Propriétaire ",sexe~"sexe"),
         missing = "no"
       ) %>% modify_header(label = "**Variable**") %>%
  bold_labels()
```

Tableau filière Riz

```
t4<-Final %>%
  dplyr::filter(filiere_4=="1") %>%
  mutate(filiere_4= recode(filiere_4, `1` = "fil anacarde")) %>%
  dplyr::select(sexe, q25, q12, q81) %>%
  gtsummary::tbl_summary(
      label = list(q25 ~ "Niveau d'inst",q12 ~ "Statut juridique",
            q81 ~ "Propriétaire ",sexe~"sexe"),
         missing = "no"
       ) %>% modify_header(label = "**Variable**") %>% bold_labels()
```

Tableau merger

```
tbl_merge(
  list(t1,t2,t3,t4,f),
  tab_spanner = c("**Fil d'arachide**",
                  "**fil anacarde**","**Fil mangue**",
                  "**fil riz**","**Total**")
                  )%>% as_flex_table()%>%
  width(width=1)
```

	Fil d'arachide	**fil anacarde**	**Fil mangue**	**fil riz**	**Total**
Variable	**N = 108**[1]	**N = 61**[1]	**N = 89**[1]	**N = 92**[1]	TOTAL =250[1]
sexe					
Femme	93 (86%)	40 (66%)	68 (76%)	77 (84%)	191 (76%)
Homme	15 (14%)	21 (34%)	21 (24%)	15 (16%)	59 (24%)
Niveau d'inst					
Aucun niveau	43 (40%)	13 (21%)	26 (29%)	11 (12%)	79 (32%)
Niveau primaire	23 (21%)	17 (28%)	24 (27%)	26 (28%)	56 (22%)
Niveau secondaire	34 (31%)	15 (25%)	25 (28%)	32 (35%)	74 (30%)
Niveau Superieur	8 (7.4%)	16 (26%)	14 (16%)	23 (25%)	41 (16%)
Statut juridique					
[1]n (%)					

	Fil d'arachide	fil anacarde	Fil mangue	fil riz	Total
Variable	**N = 108**[1]	**N = 61**[1]	**N = 89**[1]	**N = 92**[1]	TOTAL =250[1]
Association	2 (1.9%)	3 (4.9%)		2 (2.2%)	6 (2.4%)
GIE	79 (73%)	35 (57%)	73 (82%)	77 (84%)	179 (72%)
Informel	23 (21%)	12 (20%)	5 (5.6%)	3 (3.3%)	38 (15%)
SA	2 (1.9%)	2 (3.3%)	3 (3.4%)	3 (3.3%)	7 (2.8%)
SARL	1 (0.9%)	6 (9.8%)	6 (6.7%)	5 (5.4%)	13 (5.2%)
SUARL	1 (0.9%)	3 (4.9%)	2 (2.2%)	2 (2.2%)	7 (2.8%)
Propriétaire					
Locataire	12 (11%)	7 (11%)	11 (12%)	9 (9.8%)	24 (9.6%)
Propriétaire	96 (89%)	54 (89%)	78 (88%)	83 (90%)	226 (90%)

[1] n (%)

- Globalement sur 4 dirigeants près de 3 sont des femmes ,les dirigeants des PME ont tendance tendance à avoir aucun niveau ou des niveau secondaire.
- La majorité soit près de 72% des PME sont des GIE et la minorité des SA constituant 2% de toute les PME.
- Presque toute les PME sont propriétaire de leur locaux
- Les PME exercant dans la filière arachide sont majoritaire , constituant un peu moins de la moitié des PME, celle de la filière Anacarde constitue juste 24% de la base.

Qualité des routes menant aux filieres

```
Final1 <- Final
Final1$q19 <- fct_na_value_to_level(Final$q19, "NA")

l1<-Final1 %>%
  dplyr::filter(filiere_1=="1") %>%
  dplyr::select(filiere_1,region,q19) %>%
  gtsummary::tbl_strata(
    strata = "filiere_1", .tbl_fun = function(data) {
      data %>%
        gtsummary::tbl_summary(
          label = list(region = "Région"),
          missing = "no",by = q19,percent = "row") })

l2<-Final1 %>%
  dplyr::filter(filiere_2=="1") %>%
  dplyr::select(filiere_2,region,q19) %>%
  gtsummary::tbl_strata(
    strata = "filiere_2", .tbl_fun = function(data) {
      data %>%gtsummary::tbl_summary(
          label = list(q1 = "Région"),
          missing = "no",by = q19,percent = "row") })

l3<-Final1 %>%
```

```
  dplyr::filter(filiere_3=="1") %>%
  dplyr::select(filiere_3,region,q19) %>%
  gtsummary::tbl_strata(
    strata = "filiere_3", .tbl_fun = function(data) {
      data %>%gtsummary::tbl_summary(
           label = list(q1 = "Région"),
          missing = "no",by = q19,percent = "row") })
l4<-Final1 %>%
  dplyr::filter(filiere_4=="1") %>%
 dplyr::select(filiere_4,region,q19) %>%
  gtsummary::tbl_strata(
    strata = "filiere_4", .tbl_fun = function(data) {
      data %>% gtsummary::tbl_summary(
          label = list(q1 = "Région"),
          missing = "no",by = q19,percent = "row") })
## Regrouper les tableau
tbl_stack(
  list(l1,l2,l3,l4),
 group_header = c("arachide",
                  "anacarde","mangue",
                  " riz")
                  )%>%
  bold_labels() %>%
  modify_header(label = "**Région**")%>%
  as_flex_table() %>%
  width(width=1)
```

		1			
Group	**Région**	**Bon état**, N = 0[1]	**Etat moyen**, N = 38[1]	**Mauvais état**, N = 29[1]	**NA**, N = 41[1]
arachide	Région				
	Diourbel	0 (0%)	15 (45%)	14 (42%)	4 (12%)
	Fatick	0 (0%)	1 (8.3%)	9 (75%)	2 (17%)
	Kaffrine	0 (0%)	6 (75%)	1 (13%)	1 (13%)
	Kaolack	0 (0%)	5 (25%)	4 (20%)	11 (55%)
	Kolda	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Saint-Louis	0 (0%)	0 (0%)	0 (0%)	1 (100%)
	Thiès	0 (0%)	10 (37%)	0 (0%)	17 (63%)
	Ziguinchor	0 (0%)	0 (0%)	1 (17%)	5 (83%)
anacarde	region				
	Dakar	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Fatick	0 (0%)	8 (38%)	8 (38%)	5 (24%)
	Kolda	0 (0%)	3 (60%)	0 (0%)	2 (40%)
	Sédhiou	0 (0%)	1 (33%)	0 (0%)	2 (67%)

[1] n (%)

		1			
Group	**Région**	**Bon état**, N = 0[1]	**Etat moyen**, N = 38[1]	**Mauvais état**, N = 29[1]	**NA**, N = 41[1]
	Ziguinchor	1 (3.2%)	5 (16%)	7 (23%)	18 (58%)
mangue	region				
	Diourbel	0 (0%)	0 (0%)	0 (0%)	1 (100%)
	Fatick	0 (0%)	1 (33%)	2 (67%)	0 (0%)
	Kaffrine	0 (0%)	4 (80%)	0 (0%)	1 (20%)
	Kaolack	0 (0%)	2 (29%)	1 (14%)	4 (57%)
	Saint-Louis	0 (0%)	4 (9.5%)	10 (24%)	28 (67%)
	Thiès	0 (0%)	9 (36%)	0 (0%)	16 (64%)
	Ziguinchor	0 (0%)	0 (0%)	1 (17%)	5 (83%)
riz	region				
	Dakar	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Fatick	0 (0%)	0 (0%)	4 (100%)	0 (0%)
	Kaffrine	0 (0%)	1 (100%)	0 (0%)	0 (0%)
	Kaolack	0 (0%)	2 (50%)	1 (25%)	1 (25%)
	Kolda	0 (0%)	2 (50%)	0 (0%)	2 (50%)
	Sédhiou	0 (0%)	1 (33%)	1 (33%)	1 (33%)
	Thiès	0 (0%)	12 (38%)	1 (3.1%)	19 (59%)
	Ziguinchor	2 (4.7%)	8 (19%)	8 (19%)	25 (58%)

[1] n (%)

- D'après le repondants,les routes menant aux filière d'arachide aucune n'est en bon état,un peu plus de la moitié des routes soont en états moyen, tandis que le reste est en mauvais état.En terme de proportion, les enterprises de Kaffrine ont le plus de routes en état moyen

Type de contrat de bail selon le statut Juridique de la PME

```
Final %>%
select(q12,q81)%>% gtsummary::tbl_summary(   ## selection des variables à utiliser
          label = list(q12 = "Statut juridique"),
          missing = "no",
          by = q81, percent = "row") %>%
  bold_labels()%>% modify_header(label = "**Statut juridique**")  %>%
  as_flex_table()
```

Statut juridique	**Locataire**, N = 24[1]	**Propriétaire**, N = 226[1]
Statut juridique		
[1] n (%)		

Statut juridique	**Locataire**, N = 24[1]	**Propriétaire**, N = 226[1]
Association	0 (0%)	6 (100%)
GIE	17 (9.5%)	162 (91%)
Informel	0 (0%)	38 (100%)
SA	1 (14%)	6 (86%)
SARL	3 (23%)	10 (77%)
SUARL	3 (43%)	4 (57%)

[1] n (%)

Ici les pourcentages sont données en lignes, d'après les resultat , peu importe le statut juridique de l'entreprise,plus de 57% d'entre eux sont propriétaires. On constate également que aucune association ne loue pendant que près de 43% des SUARL sont locataires.

3 Un peu de cartographie

3.1 Transformer le data.frame en données géographiques dont l'objet sera nommé projet_map.

Il est nécessaire d'importer la bibliothèque sf

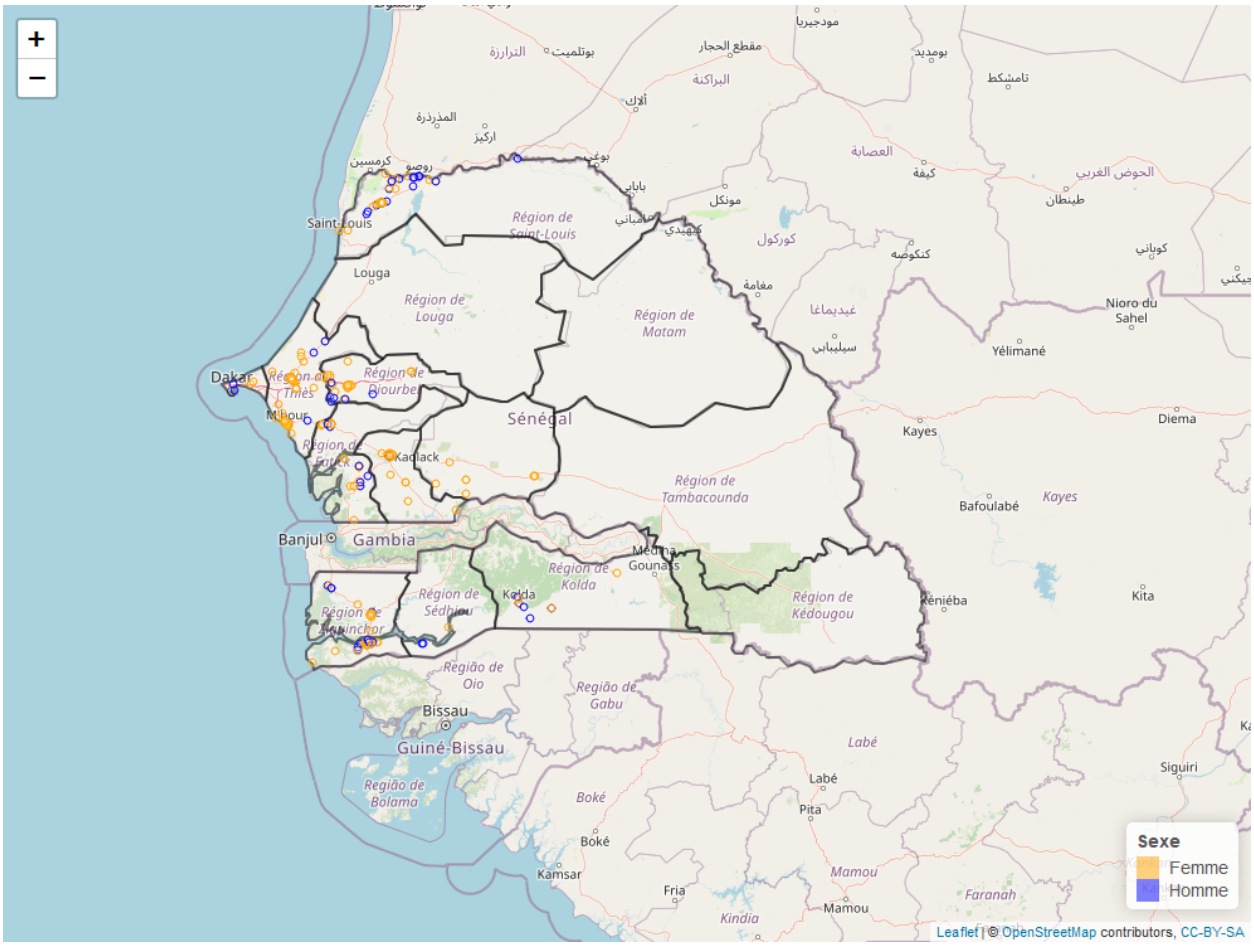
```
projet_map <- st_as_sf(projet, coords = c("gps_menlongitude", "gps_menlatitude"), crs = 4326)
```

3.2 Faites une réprésentation spatiale des PME suivant le sexe

Ici nous allons utiliser la librairies rnaturalearth pour avoir les données polygones du Sénégal, ensuite la librairie leaflet qui nous permettra de faire les cartes .Ausquels nous ajouterons des légendes en fonction de l'analyse rechercher. Etant données que leaflet a des cartes interactives nous allons creer une capture d'écran que nous allons joindre à notre travail, il sera nécessaire d'importer les librairies webshot et htmlwidgets.

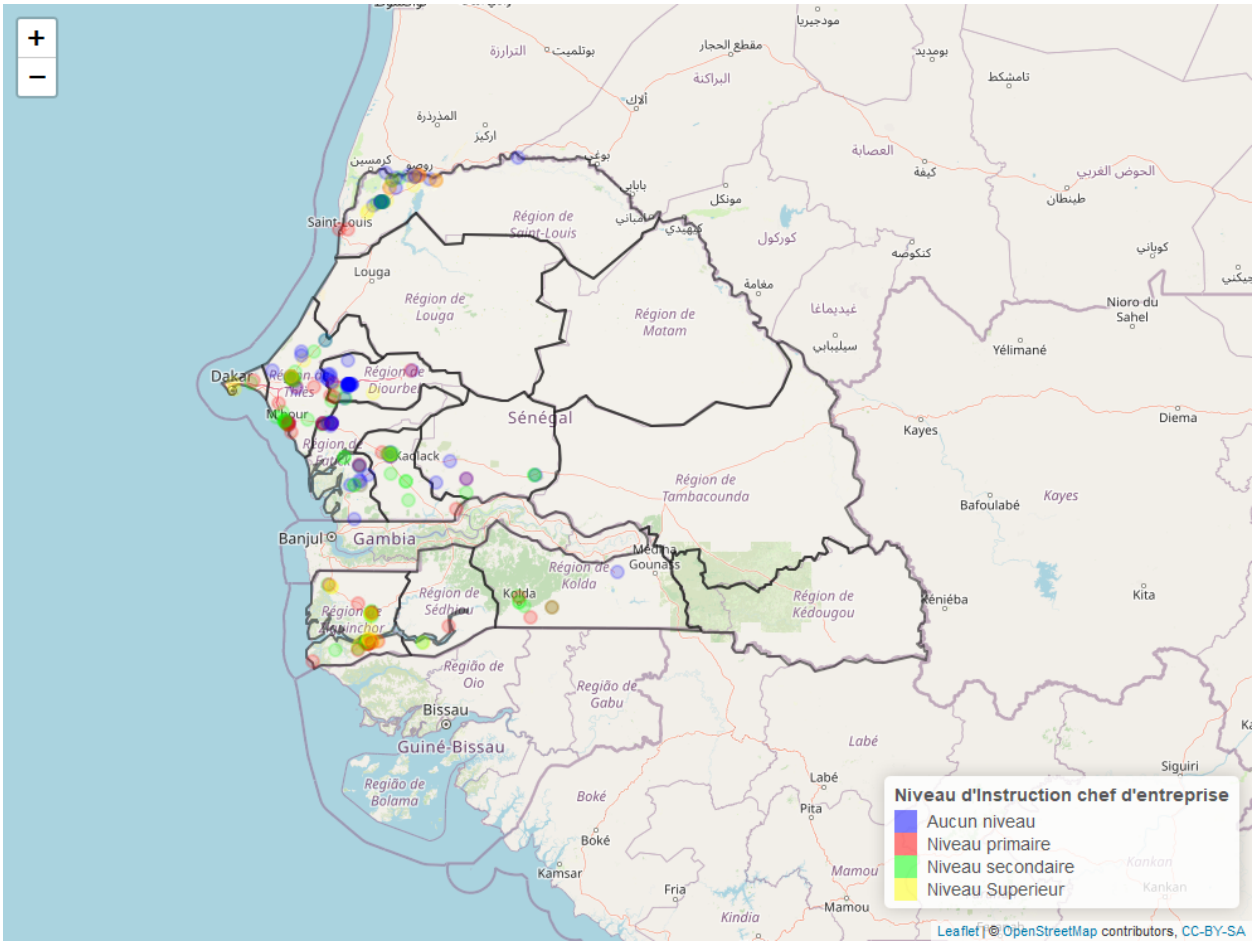
```
## importation des donées du Sénégal
Senegal <- rnaturalearth::ne_states( country = "Senegal",returnclass = "sf")

## Vecteurs de couleurs de légende en fonction du sexe
couleurs_sexe <- colorFactor(c("orange", "blue"), domain = projet_map$sexe)

# creation de la carte

v<-leaflet() %>%
    setView(lng = -14, lat = 14, zoom = 7) %>%
    addTiles() %>% #Ligne à désactiver si l'on veut seulement afficher la carte de l'Afrique de l'ouest
    addPolygons(
      data = Senegal,
      fillColor = "white",
      color = "black",
      weight = 2,
      fillOpacity = 0.1
    )%>%
    addCircleMarkers(        ##markers sous forme de cercle
      lng = projet$"gps_menlongitude",
      lat = projet$"gps_menlatitude",
      layerId = NULL,
```

```
      radius = 3,
      weight = 1,
      opacity = 0.2,
      fill = TRUE,
      fillColor = "white",
      fillOpacity = 0.2,
      color = ifelse(projet$sexe=="Femme","orange","blue"),
      popup = ifelse(projet$sexe=="Femme","femme","homme"),
      label = ifelse(projet$sexe=="Femme","femme","homme"),
      options = markerOptions(),
      data = projet_map
    ) %>%
  addLegend(position = "bottomright",     # placer la légende à  droite au fond
  pal = couleurs_sexe,
  values = projet_map$sexe, title = "Sexe")


## CApture
saveWidget(v, file = "Sexe.html")
webshot("Sexe.html", "Sexe.png")
```



1.3.3. Répartition des PME suivant le niveau d'instruction

```
# Définir les couleurs pour chaque valeur de la colonne "niv d'instruction"
couleurs_niv <- colorFactor(c("blue", "red","green","yellow"), domain = projet_map$q25)



t <- leaflet() %>%
   setView(lng = -14, lat = 14, zoom = 7) %>%
  addTiles() %>%
  addPolygons(
      data = Senegal,
      fillColor = "white",
      color = "black",
      weight = 2,
      fillOpacity = 0.1
      ) %>%
  addCircleMarkers(data = projet_map,color = ~couleurs_niv(q25),
      lng = projet$"gps_menlongitude",
      lat = projet$"gps_menlatitude",
      layerId = NULL,
      radius = 5,
      weight = 2,
      opacity = 0.2,
      fill = TRUE,
      fillOpacity = 0.2,)%>%
  addLegend(position = "bottomright",
  pal = couleurs_niv,
  values = projet_map$q25, title = "Niveau d'Instruction chef d'entreprise")


#Nous  allons faire une capture d'écran de celle-ci avec la fonction webshot.


saveWidget(t, file = "Niv.html")
webshot("Niv.html", "Niv.png")
```



Repartion des PME par Région

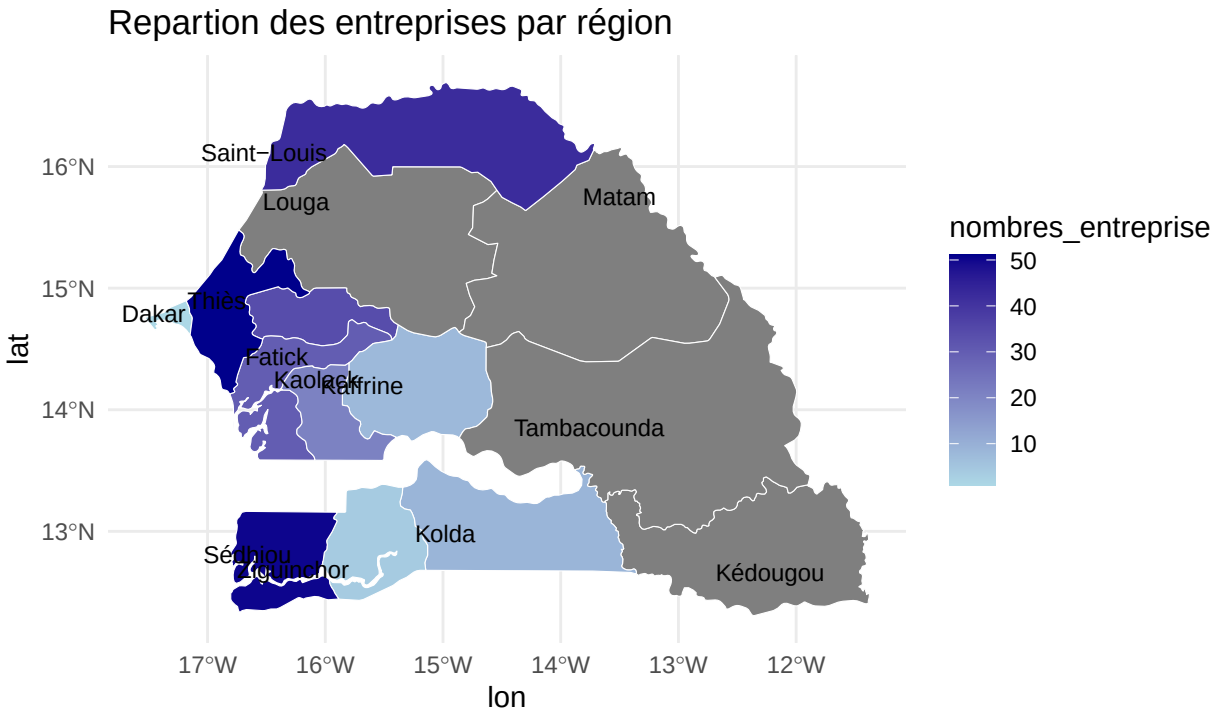
Création de du dataframe qui compte les entreprises par région.

```
#projet$nombre <- projet %>%dplyr::mutate(nombre==1)

projeti <- projet %>%
  group_by(region) %>%
  count() %>%ungroup()
projeti<- projeti %>%mutate(nombres_entreprise=n)
```

Création de la carte

```
Senegal1<-merge(Senegal,projeti,by.x = "name", by.y = "region", all.x = TRUE)

## AJouter les coordonnées des régions
regions <- data.frame(
  region = c("Dakar", "Thiès", "Fatick", "Kaolack", "Kaffrine", "Kédougou", "Kolda", "Louga", "Matam",
  lon = c(-17.455390, -16.920337, -16.412964, -16.073365, -15.687128, -12.220533, -14.981073, -16.24638
  lat = c(14.693425, 14.798658, 14.339950, 14.151858, 14.101164, 12.559404, 12.887101, 15.614472, 15.65
)

#senegal_centroid <- st_centroid(Senegal1)
# Créer la carte
ggplot(data = Senegal1) +
```

```
  geom_sf(aes(fill = nombres_entreprise), color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  theme_minimal()+
  geom_text(data=regions,aes(x=lon,y=lat,label=region),size=3,nudge_y=0.1)+
  ggtitle("Repartion des entreprises par région")
```



PARTIE 2

Importation de la base

```
projet2<- read_excel("Base_Partie 2.xlsx",
                     sheet = "data",
                     range = NULL,
                     col_names = TRUE,
                     col_types = NULL)
```

##2. 1 Nettoyage et gestion des données

2.1.1 Renommer la variable “country_destination” en “destination” et définir les valeurs négatives comme manquantes.

```
projet2 <- projet2 %>%
  mutate(destination = case_when(
    projet2$country_destination < 0 ~ NA,
    TRUE ~ projet2$country_destination
```

```
  ))
```

2.1.2. Création d'une nouvelle variable contenant des tranches d'âge de 5 ans en utilisant la variable "age"

Après exploration de la base, nous avons vu un âge dun age de 999, qui est absurde que nous allons par la suite imputer à la moyenne des âges sans cette observation.

```
projet2$age [projet2$age == 999] <- round(
  mean(projet2$age[projet2$age != 999], na.rm = TRUE)
)

# Création des classes

inter <- 5
limites_classes <- seq(min(projet2$age), max(projet2$age), by = inter)

# Créationr des classes d'âge en utilisant cut()

projet2$classes_age <- cut(projet2$age, breaks =
      limites_classes, labels = paste0("{",limites_classes
          [-length(limites_classes)], ";", limites_classes[-1],"{"))

## placer la nouvelle variable creer pres de sexe
projet2 <- projet2%>% relocate(classes_age, .after = sex)

kable(projet2[1:3,3:8],format="latex")
```

endtime	enumerator	district	age	sex	classes_age
2019-01-14 15:11:10	6	1	33	1	{30;35{
2019-01-14 16:45:52	6	1	43	0	NA
2019-01-14 17:45:47	6	1	28	0	{25;30{

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

```
projet2 <- projet2 %>%
  group_by(enumerator) %>%
  mutate(nombre_entretiens = n()) %>%ungroup()

## relocalisation de la variable
projet2 <- projet2%>% relocate(nombre_entretiens, .after = enumerator)
kable(projet2[1:3,1:6], format="latex")
```

id	starttime	endtime	enumerator	nombre_entretiens	district
2	2019-01-14 14:56:37	2019-01-14 15:11:10	6	5	1
3	2019-01-14 16:12:22	2019-01-14 16:45:52	6	5	1
4	2019-01-14 17:15:47	2019-01-14 17:45:47	6	5	1

2.1.4. Création d'une nouvelle variable qui affecte aléatoirement chaque répondant à un groupe de traitement (1) ou de controle (0)

```
set.seed(124)
projet2 <- projet2 %>%
```

```
  mutate(groupe = sample(c(0, 1), size = nrow(projet2),
                         replace = TRUE))
```

2.1.5. Fusioner les base district et data

```
feuille_2 <- read_excel("Base_Partie 2.xlsx",
    sheet = "district")

projet2 <-merge(projet2 , feuille_2,by="district")
```

2.1.6. Durée et Durée moyenne de l'entretien

Determinons tout d'abord les durées les entretiens par interview, nous allons ensuite les regrouper par enquêteur afin de calculer la moyenne des durée des enquêtes par enquêteurs.Nous allon importer la librairies lubridate qi permet de faire la manipulation sur les variables temporelles.

```
projet2 <-projet2 %>%
  mutate(duree_entretien=endtime-starttime )

## RELOCALISATION
projet2 <- projet2%>% relocate(duree_entretien, .after = nombre_entretiens)

# MOYENNE par enqueteur
projet2 <- projet2 %>%
  group_by(enumerator) %>%
  mutate(duree_moyen =sum(duree_entretien)/nombre_entretiens) %>%ungroup()

# Relocalisation
projet2 <- projet2%>% relocate(duree_moyen, .after = duree_entretien)

kable(projet2[1:3,4:8], format="latex")
```

endtime	enumerator	nombre_entretiens	duree_entretien	duree_moyen
2019-01-14 15:11:10	6	5	14.55 mins	25.84667 mins
2019-01-14 16:45:52	6	5	33.50 mins	25.84667 mins
2019-01-14 17:45:47	6	5	30.00 mins	25.84667 mins

2.1.7. Renommage des variables en y ajoutant le suffixe endline

```
projet3 <- projet2 %>% rename_all(~ paste0("endline", .))

kable(projet3[1:3,1:3], format="latex")
```

endlinedistrict	endlineid	endlinestarttime
1	2	2019-01-14 14:56:37
1	3	2019-01-14 16:12:22
1	4	2019-01-14 17:15:47

2.2. Analyse et visualisation des données

2.2.1. Tableau récapitulatif de l'age moyen et d'enfants moyen par district

```
mean_tab <- flextable::as_flextable(projet2 %>%
          group_by(district) %>% summarise(Age_Moyen = mean(age),
```

```
          Enfant_Moyen = mean(children_num)))
mean_tab
```

district	Age_Moyen	Enfant_Moyen
numeric	numeric	numeric
1	29.6	1.5
2	26.6	0.9
3	26.1	0.0
4	26.0	0.0
5	24.3	0.5
6	23.2	0.1
7	28.0	0.2
8	24.6	1.3

2.2.2. Testons si la différence d'âge entre les sexes est statistiquement significative au niveau de 5 %

```
test<-projet2 %>%
dplyr::select(sex ,age) %>%
gtsummary::tbl_summary(
  by = sex,
  label = age ~ "Tranche d'age",
  percent = "column",
                    statistic=age~"{mean}")%>% add_n() %>%
  add_difference() %>%
  bold_labels() %>%
   as_flex_table()%>%
  width(width=1)
test
```

Characteristic	**N**	**0**, N = 86[1]	**1**, N = 11[1]	**Difference**[2]	**95% CI**[23]	**p-value**[2]
Tranche d'age	97	26	22	3.6	-0.02, 7.3	0.051

[1]Mean

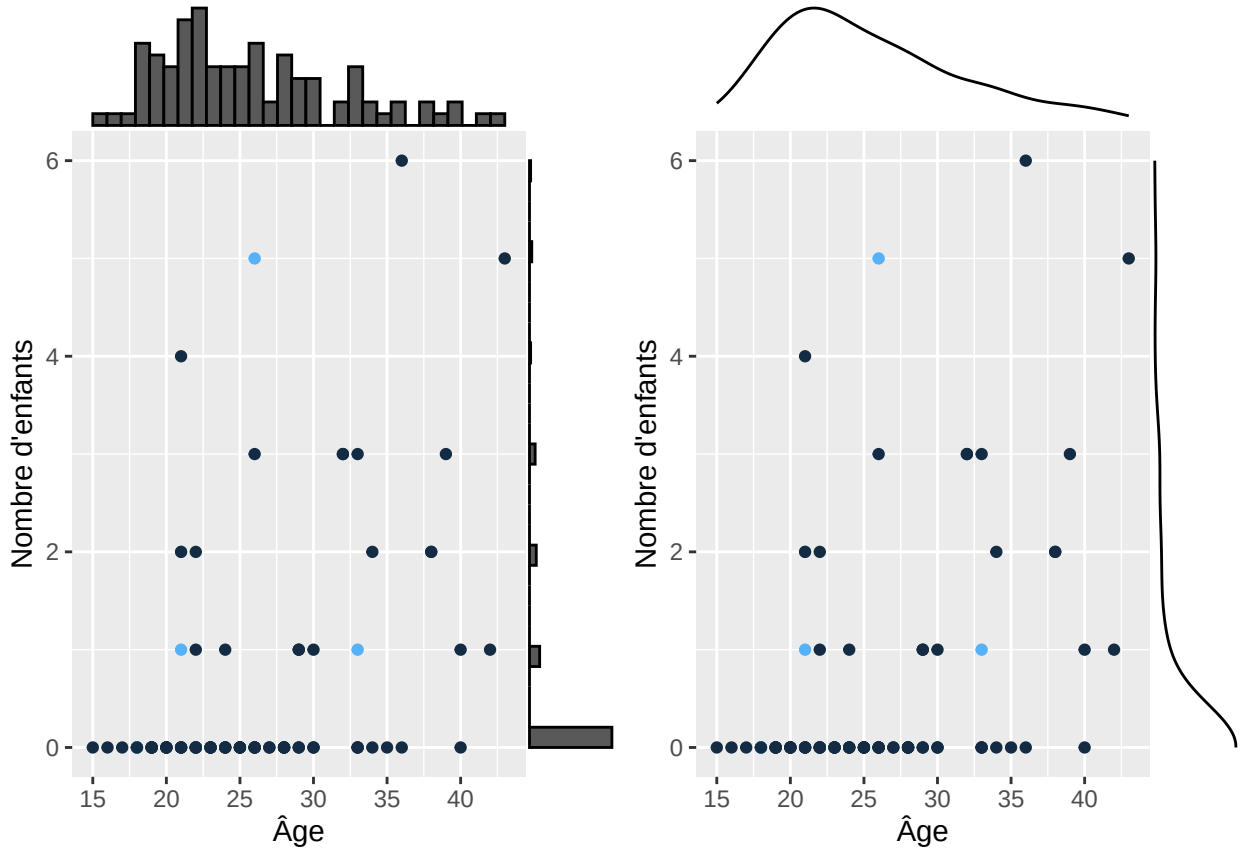
[2]Welch Two Sample t-test

[3]CI = Confidence Interval

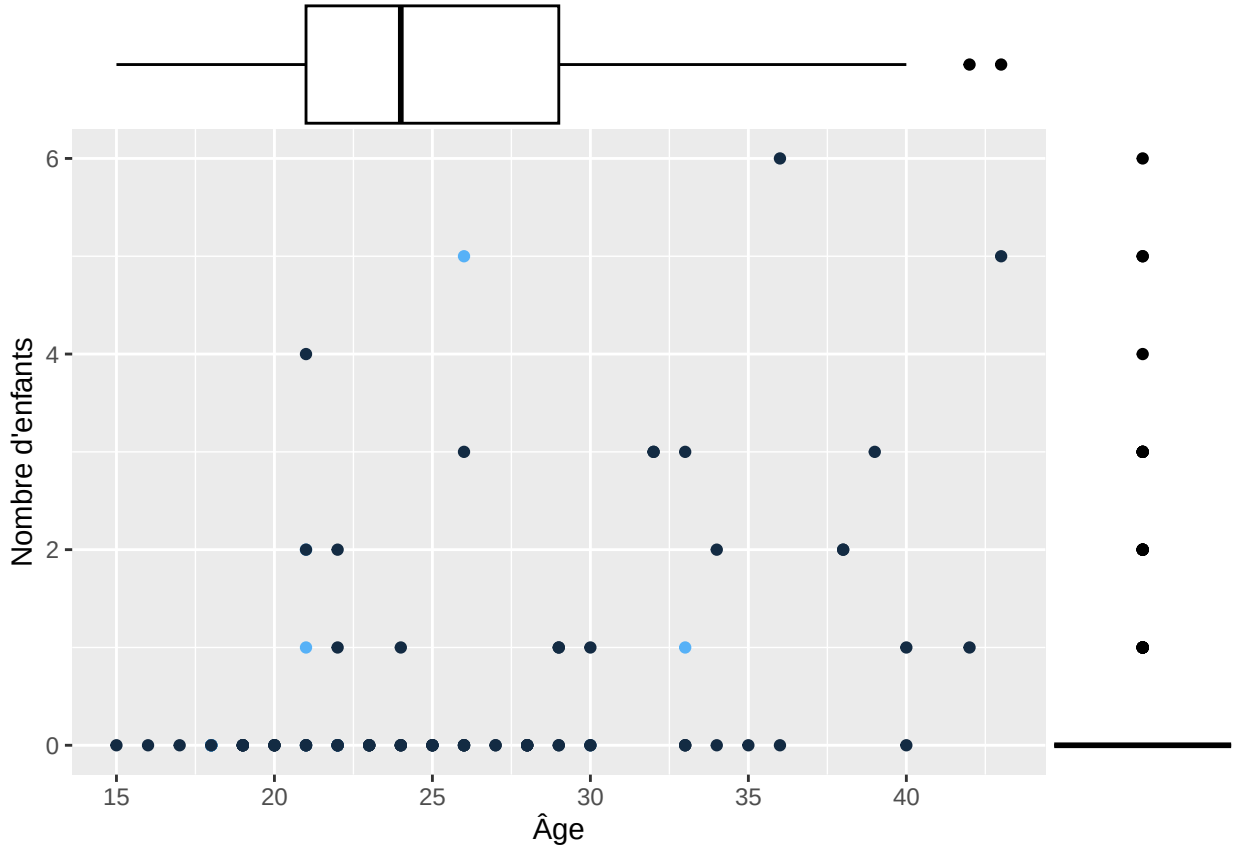
2.2.3. Nuage de points de l'âge en fonction du nombre d'enfants

On importe les librairies ggplot et ggExtra pour faire nos differents graphes

```r
graph <- ggplot(projet2, aes(x=age, y=children_num, color=sex)) +
      geom_point() +
      theme(legend.position="none")+
      labs(x = "Âge", y = "Nombre d'enfants")

graph1 <- ggMarginal(graph, type="histogram")
graph2 <- ggMarginal(graph, type="density")
graph3 <- ggMarginal(graph, type="boxplot")


## POsitionner les différents  graphes dans la grille
grid.arrange(graph1, graph2,ncol = 2)
```




```r
grid.arrange(graph3,ncol = 1)
```



- Ici nous avons fait sortir le nuage de point , l'histogramme et la densité des distributions

2.2.4. Estimation de l'effet de l'appartenance au groupe sur la décision de migrer

Nous allons importer la librairies nnet, GGally,effects, gridExtra qui permettre de faire des regressions. Nous allons utiliser la variable de traitement ou controle, groupe que nous avons creer plus haut.

```
regm <- multinom(intention ~ groupe ,data = projet2)
```

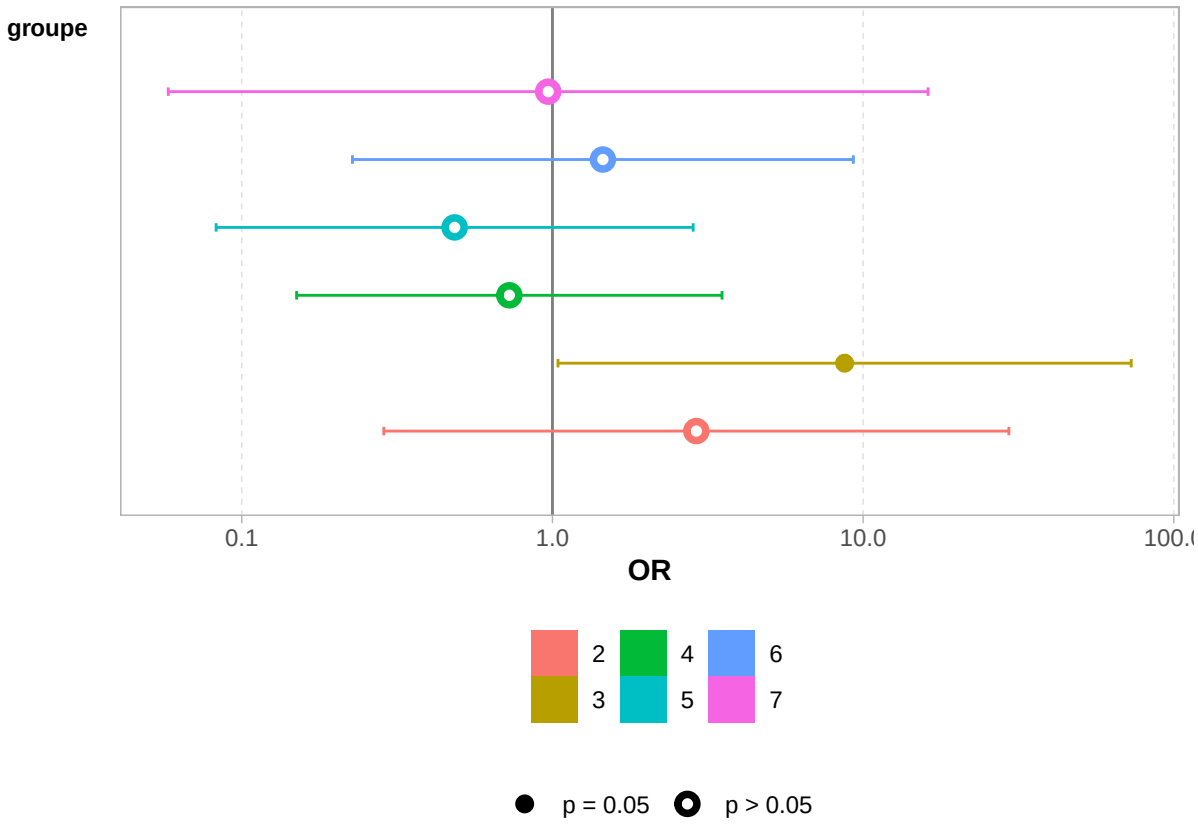
```
## # weights:  21 (12 variable)
## initial  value 188.753284
## iter  10 value 116.054405
## final  value 116.047188
## converged
```

```
tbl1 <- tbl_regression(regm, exponentiate = TRUE)
tbl1
```

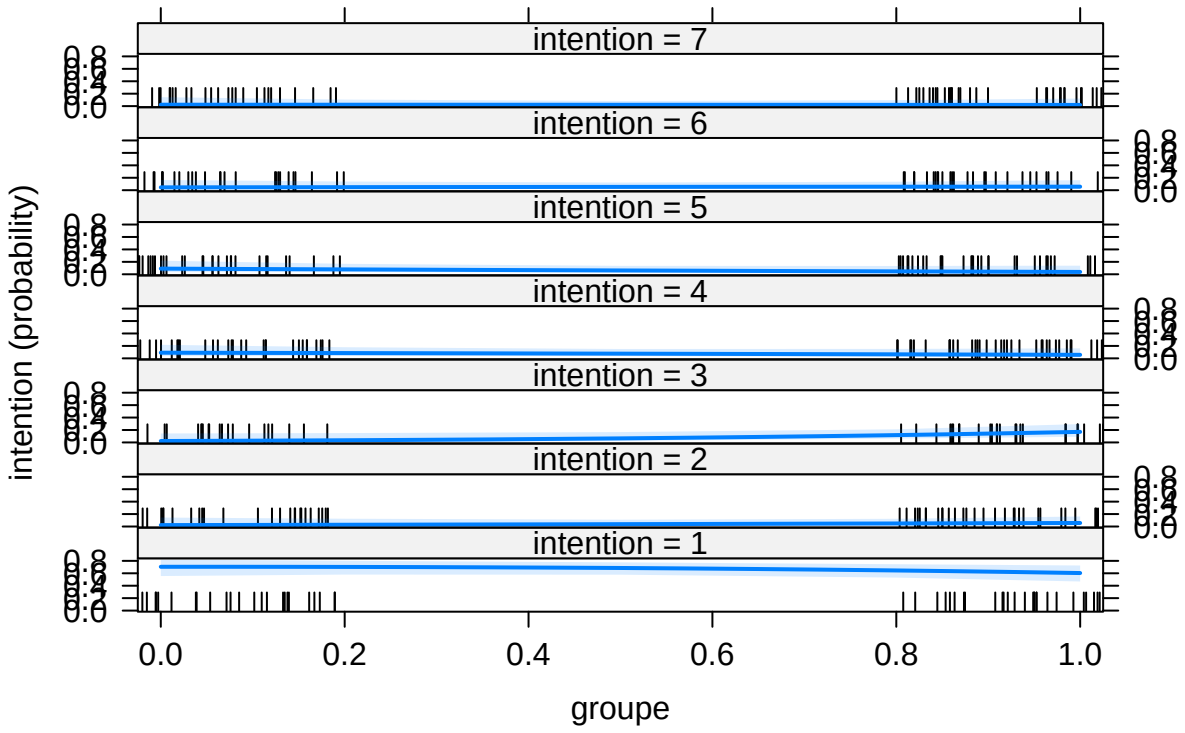
Outcome	**Characteristic**	**OR**	**95% CI**	**p-value**
2	groupe	2.91	0.29, 29.5	0.4
3	groupe	8.72	1.04, 72.9	0.046
4	groupe	0.73	0.15, 3.51	0.7
5	groupe	0.48	0.08, 2.84	0.4
6	groupe	1.45	0.23, 9.30	0.7
7	groupe	0.97	0.06, 16.2	>0.9

```
# Visualisation du modèle créer
ggcoef_multinom(
  regm,
```

```
  exponentiate = TRUE
)
```




```
# visualisons l'effet marginal des variables
plot(allEffects(regm))
```



2.2.5. Tableau de regression avec trois modèles

```
modele <- stats::lm(intention ~ groupe, data = projet2)
modele %>% gtsummary::tbl_regression()
```

Characteristic	**Beta**	**95% CI**	**p-value**
groupe	0.05	-0.66, 0.75	0.9

```
# Modèle A : Modèle vide - Effet du traitement sur les intentions
modele_A <- stats::lm(intention ~ groupe, data = projet2)
tableau_A <- tbl_regression(modele_A)

# Modèle B : Effet du traitement sur les intentions en tenant compte de l'âge et du sexe
modele_B <- stats::lm(intention ~ groupe + age + sex, data = projet2)
tableau_B <- tbl_regression(modele_B)

# Modèle C : Identique au modèle B mais en contrôlant le district
modele_C <- stats::lm(intention ~ groupe + age + sex + district, data = projet2)

tableau_C <- tbl_regression(modele_C)

# Création du tableau récapitulatif des résultats des trois modèles
tableau_final <- tbl_merge(list(tableau_A, tableau_B, tableau_C),
              tab_spanner = c("Modèle A","Modèle B","Modèle C"))%>% as_flex_table()%>%
  fontsize() %>%
  width(width=1)
```

```
# affichage
tableau_final
```

	Modèle A			Modèle B			
Characteristic	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**	**Be**
groupe	0.05	-0.66, 0.75	0.9	0.12	-0.58, 0.83	0.7	0.(
age				0.00	-0.06, 0.06	>0.9	0.(
sex				-0.95	-2.1, 0.17	0.10	-0.
district							0.(

[1]CI = Confidence Interval

PARTIE 3

Dans cette section j'ai creer un fichier de code Ivana_KUETE_app pour l'application et j'ai uniquement utiliser les données de la base ACELD;CVS