REPUBLIQUE DU SENEGAL

Un Peuple-Un But-Une Foi

Ministère de l'Economie du Plan et de la Coopération

ANSD
Agence Nationale de la Statistique et de la Démographie

ensae

PROJET R : Première partie

24 JUILLET 2023

By KABASSINA Gnimdou Ange
Elève Ingénieure Statisticienne Economiste

Sommaire

Importation et mise en forme

Importation de la base de données

```
# chargement du package nécessaire à l'importation
library(readxl)

# Importation de la base sous format data.frame da,s l'objet projet
projet = readxl::read_excel("Base_Partie 1.xlsx")
```

Selection des variables

Ici il est question de selectionner toutes les variables à l'éxception de la première "key".

```
# chargement du package nécessaire au traitement de la base
library(dplyr)

# selection de toutes les variables sauf "key"
# Importons `magrittr` pour acceder aux pipes
library(magrittr)
projet_select = projet %>% dplyr::select(2:last_col())
```

Vérification des valeurs manquantes

Tabulation du nombre de valeurs manquantes par variables

Pour ce faire, nous allons chercher le nombre de valeurs maquantes pour chaque variables grâce à la commande "sum(is.na())" que nous allons appliquer sur toutes les colones de notre dataframe grâce à une boucle for comme suit :

```
## liste pour contenir le nombre de variables manquantes
var_na = c()
for (i in 2:length(projet)){
  var_na = c(var_na, projet %>% dplyr::select(i) %>% is.na() %>% sum())
}
```

Une fois le nombre de valeurs manquantes par variables obtenues, nous allons créer un dataframe pour le contenir puis transformer ce dernier en tableau de format markdown (avec les tirets qui seviendra un vrai tableau après avoir été généré).

```
## stocker le résultat (variables + nombre de na) dans un dataframe
tab_na <- data.frame(variable = names(projet)[2:length(projet)], nombre_na = var_na)
## ajouter une colone avec la proportion de NA:
tab_na <- tab_na %>% mutate(proportion_na = nombre_na*100/nrow(projet))

## Afficher le data.frame sous forme de tableau au format Markdown
### nous allons utiliser kable_styling du package"kableExtra" pour mettre en forme notre tableau
knitr::kable(tab_na, format = "markdown")
```

variable	nombre_na	proportion_na
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

Vérification des valeurs manquantes pour la variable “key”

A présent nous allons prendre un cas particulier, celui de la variable “key” et vérifier si elle présente des valeurs manquantes toujours en utilisant sum(is.na())

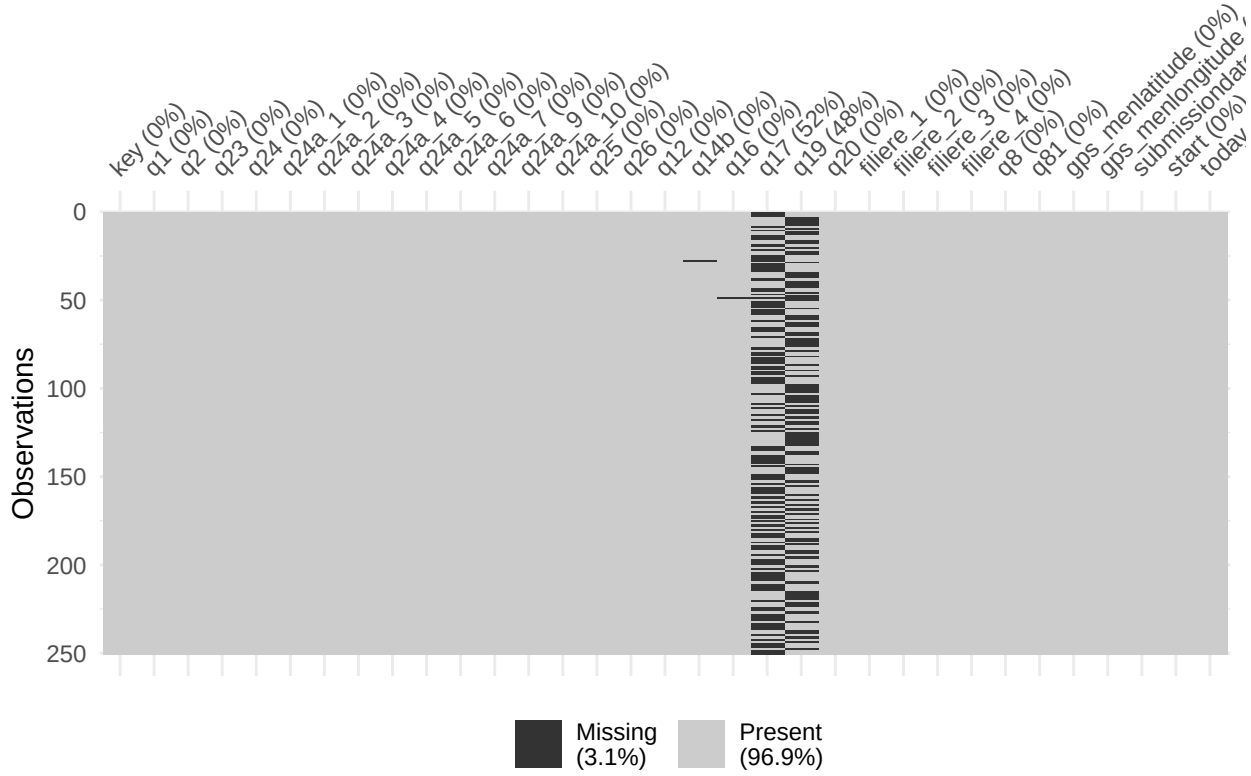
```
# vefication de l'existance de valeurs manquantes pour la variable "key"
projet %>% dplyr::select("key") %>% is.na() %>% sum()
```

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

Il n’y a donc pas de valeurs manquantes pour la variable “key”.

Essayons à présent de vusualiser le taux de valeurs manquantes dans l’ensemble de la base, par variable de même que leur position grâce à la fonction vis_miss() du package visdat

```
#visualisation avec vis_miss
library(visdat)
visdat::vis_miss(projet)
```



Création de variables

Renomination des variables q1, q2 et q23

```
# renomination de variables
projet = projet %>% dplyr::rename("region" = "q1", "departement" = "q2", "sexe"="q23")
```

Création de la variable “sexe_2”

```
#création de la variable "sexe_2"
projet = projet %>% dplyr::mutate(sexe_2 = ifelse(sexe=="Femme",1,0))
```

Création du data.frame langue

Nous allons pour créer le data.frame langue, selectionner la variable key de même que toutes les variables portant sur les langues c’est à dire toutes les variables commençant par “q24a_”.

```
# Création du data.frame langue
langue = projet %>% dplyr::select("key",starts_with("q24a_"))
```

Création de la variable parle

La variable parle devant représenter le nombre de langue parler par le dirigeant de la PME sera calculé en faisant la somme des variables relatives aux langues puisque ces dernière prenne la valeur 1 quand le dirigeant parle la langue en question et 0 sinon.

```
# Création de la variable parle = nombre de langgues parlées
langue = langue %>% dplyr::mutate(parle = rowSums(dplyr::select(langue,starts_with("q24a_"))))
```

Selection des variables les variables key et parle

```
# Selection des variables les variables key et parle
langues = langue %>% dplyr::select("key","parle")
```

Fusion des data.frame projet et langues

```
# fusion des data.frame projet et langues
projet = merge(projet,langues, by = "key")
```

Analyses descriptives

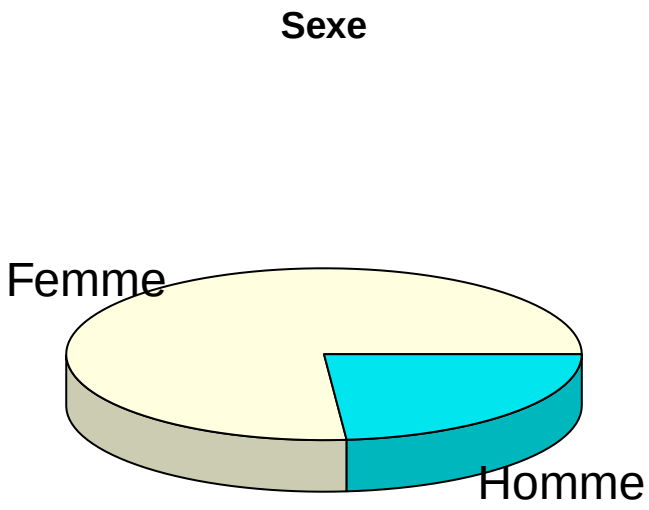
Répartition des PME suivant le sexe du dirigeant

L'analyse du graphique ci-dessous révèle que plus de 7 PME sur 10 étudiées soit 76.4 % sont dirigées par des femmes.

```
# Répartition des PME suivant le sexe du dirigeant
# Tableau de fréquences
knitr::kable(prop.table(table(projet$sexe))* 100, format = "markdown")
```

Var1	Freq
Femme	76.4
Homme	23.6

```
# graphique en secteur
library(plotrix)
plotrix::pie3D(table(projet$sexe),labels = c("Femme","Homme") ,col = c("lightyellow","turquoise2"), mai
```



Répartition des PME suivant le niveau d'instruction du dirigeant

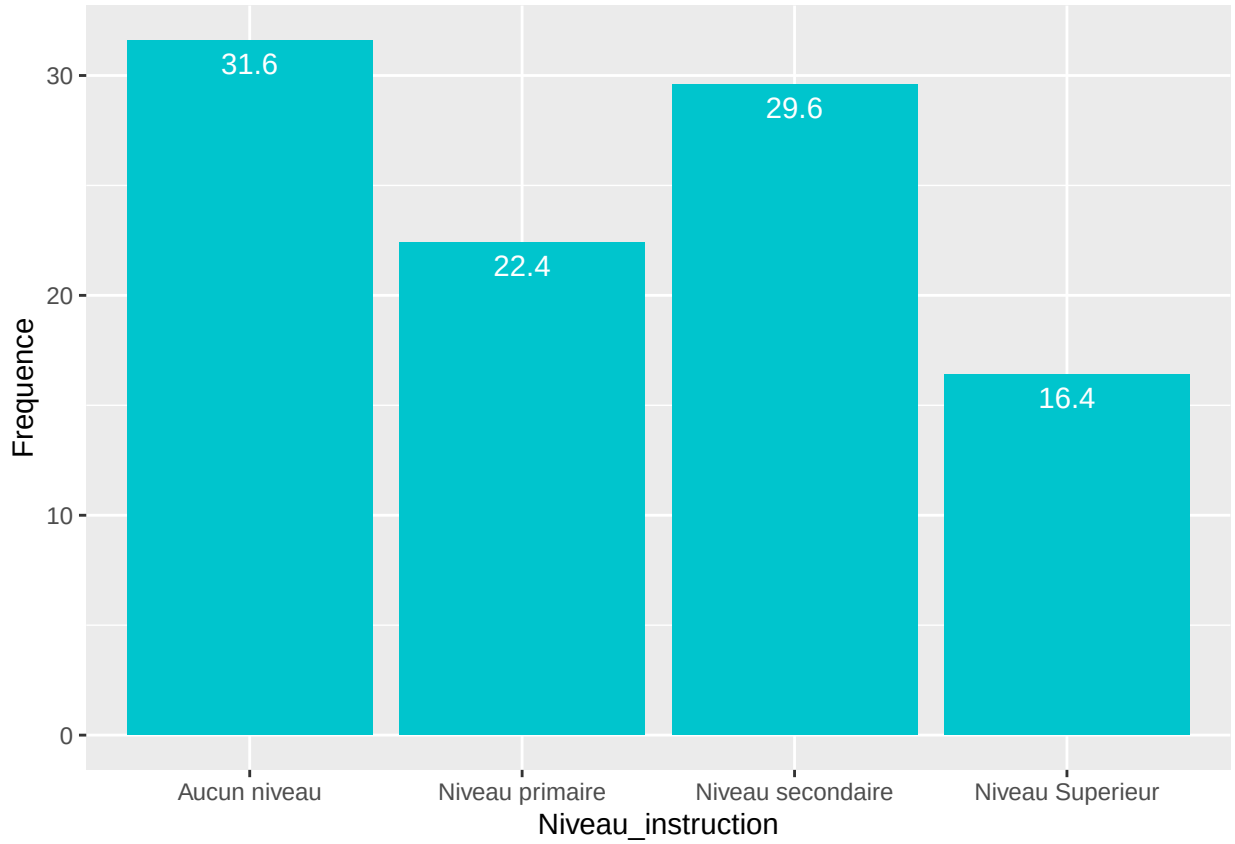
Concernant le niveau d'instruction, la majeure partie soit 31.6 % des dirigeants des PME de notre population n'ont aucun diplôme pendant que seulement 16 sur 100 (soit 16.4 %) ont fait des études supérieures.

```
# Répartition des PME suivant le niveau d'instruction du dirigeant
# Tableau de fréquences
df = data.frame(prop.table(table(projet$q25))*100) # tableau de fréquence transformé en data.frame
# renommons les colones de notre data.frame de façon plus significatives
df = df %>% dplyr::rename("Niveau_instruction" = "Var1", "Frequence" = "Freq")

# sortir le tableau sous format markdown
knitr::kable(df, format = "markdown")
```

Niveau_instruction	Frequence
Aucun niveau	31.6
Niveau primaire	22.4
Niveau secondaire	29.6
Niveau Superieur	16.4

```
### diagramme en baton
library(ggplot2)
ggplot(df, aes(x = Niveau_instruction, y = Frequence))+
```

```
  geom_col(fill = "turquoise3")+
  geom_text(aes(label = Frequence), vjust = 1.6, color = "white")
```



Répartition des PME suivant le statut juridique

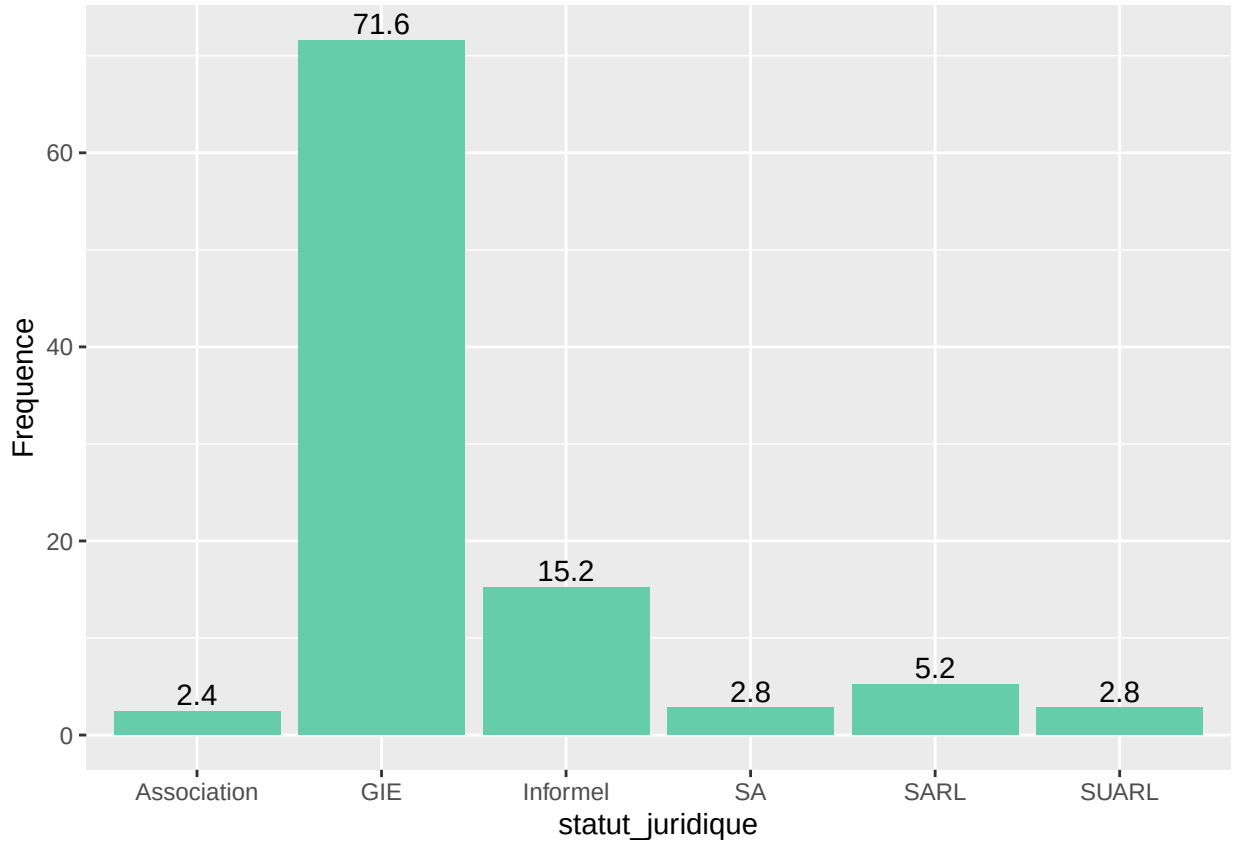
Une analyse du statut juridique des entreprises consignée dans le graphique ci-après suggère que sur les 250 PME étudiés, 179 soit 71.6 % sont des Groupement d'Intérêt économique (GIE) et que près de 15 PME sur 100 évoluent dans le secteur informel.

```
# Répartition des PME suivant le statut juridique

# Tableau de fréquences
df = data.frame(prop.table(table(projet$q12))*100) # tableau de fréquence transformé en data.frame
# renommons les colones de notre data.frame de façon plus significative
df = df %>% dplyr::rename("statut_juridique" = "Var1", "Frequence" = "Freq")

# sortir le tableau sous format markdown
knitr::kable(df, format = "markdown")
```

statut_juridique	Frequence
Association	2.4
GIE	71.6

statut_juridique	Frequence
Informel	15.2
SA	2.8
SARL	5.2
SUARL	2.8

```
### diagramme en baton
ggplot(df, aes(x = statut_juridique, y = Frequence))+
  geom_col(fill = "aquamarine3")+
  geom_text(aes(label = Frequence), vjust = -0.3, color = "black")
```



Répartition des PME suivant le statut de propriété par rapport au local de l'entreprise (propriétaire ou locataire)

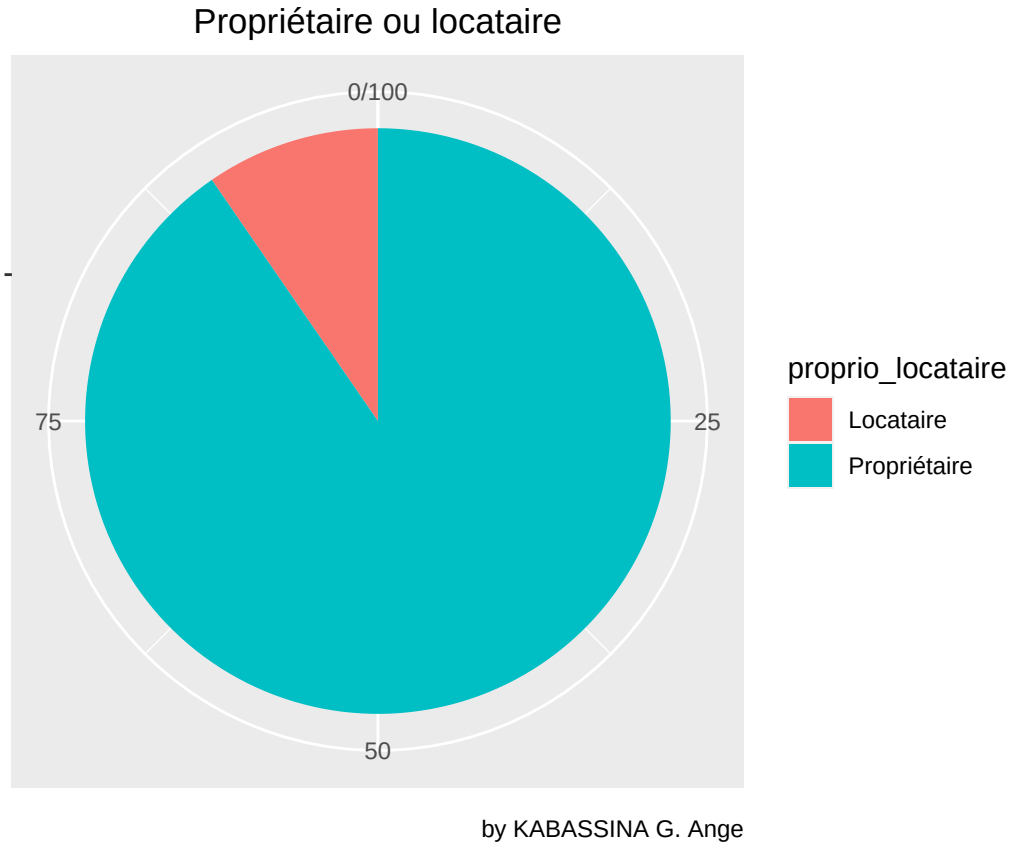
En s'interressant au local occupé par les PME, ont remarque que la quasi totalité des PME étudiée soit jusqu'à 90.4 % sont propriétaires de leur lieu de travail.

```
### Répartition des PME suivant le statut de propriété par rapport au local
### de l'entreprise (propriétaire ou locataire)*

# Tableau de fréquences
df = data.frame(prop.table(table(projet$q81))*100) # tableau de fréquence transformé en data.frame
# renommons les colones de notre data.frame de façon plus significative
```

```r
df = df %>% dplyr::rename("proprio_locataire" = "Var1", "Frequence" = "Freq")

# sortir le tableau sous format markdown
knitr::kable(df, format = "markdown")
```

proprio_locataire	Frequence
Locataire	9.6
Propriétaire	90.4

```r
# graphique en secteur
ggplot(df, aes(x = "", y=Frequence, fill = factor(proprio_locataire))) +
    geom_bar(width = 1, stat = "identity") +
    theme(axis.line = element_blank(),
          plot.title = element_text(hjust=0.5)) +
      labs(fill="proprio_locataire",
                    x=NULL,
                    y=NULL,
                    title="Propriétaire ou locataire",
                    caption="by KABASSINA G. Ange") +
  coord_polar(theta = "y", start=0)
```



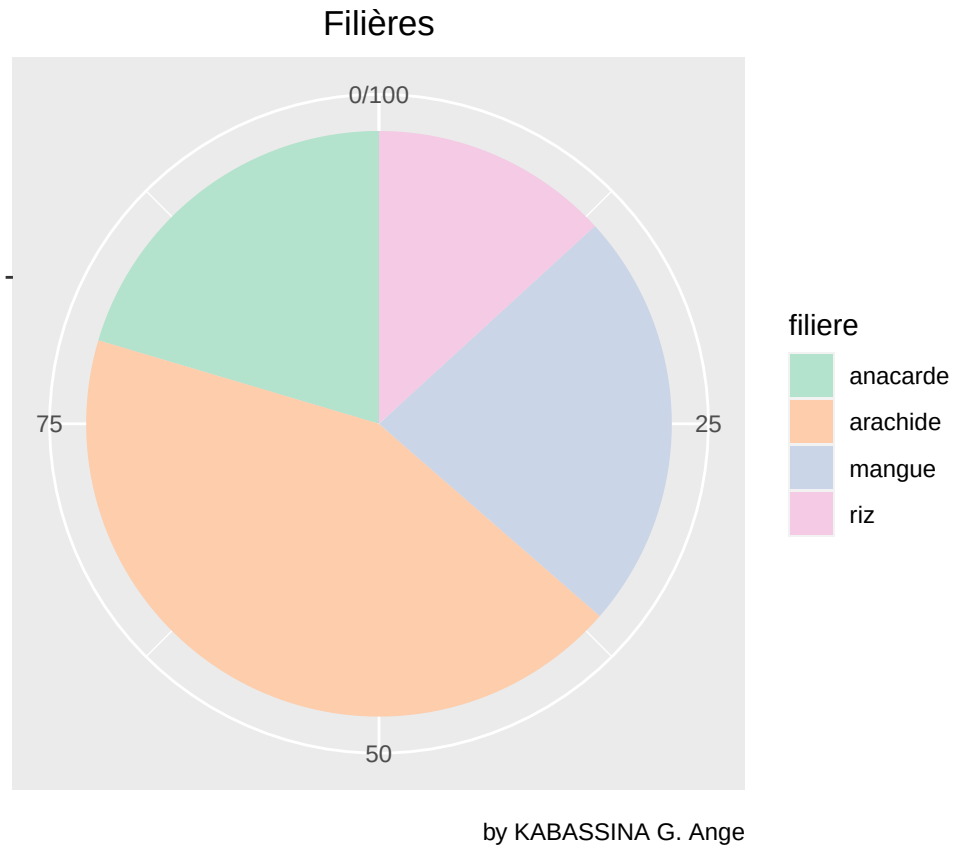
Répartition des PME suivant la filière d'activité

Pour l'analyse par filière, nous allons créer une nouvelle variable "filiere". Cette variable sera créée en fonction de toutes les variables portant sur les filières et prendra le nom de la filière x pour laquelle la valeur de la variable "est dans la filière x" est 1.

```
# Répartition des PME suivant la filière
# création de la variable filiere
projet = projet %>% dplyr::mutate(filiere = ifelse(filiere_1==1,"arachide",
                                                   ifelse(filiere_2==1,"anacarde",
                                                     ifelse(filiere_3==1,"mangue",
                                          ifelse(filiere_4==1,"riz","n_s_p")))))


# Tableau de fréquences
df = data.frame(prop.table(table(projet$filiere))*100) # tableau de fréquence transformé en data.frame
# renommons les colones de notre data.frame de façon plus significative
df = df %>% dplyr::rename("filiere" = "Var1", "Frequence" = "Freq")

# sortir le tableau sous format markdown
knitr::kable(df, format = "markdown")
```

filiere	Frequence
anacarde	20.4
arachide	43.2
mangue	23.2
riz	13.2

```
# graphique en secteur
library(RColorBrewer) #pour utiliser une palette de couleur prédéfinie

ggplot(df, aes(x = "", y=Frequence, fill = factor(filiere))) +
    geom_bar(width = 1, stat = "identity") +
  scale_fill_brewer(palette = "Pastel2") +
    theme(axis.line = element_blank(),
          plot.title = element_text(hjust=0.5)) +
       labs(fill="filiere",
                    x=NULL,
                    y=NULL,
                    title="Filières",
                    caption="by KABASSINA G. Ange") +
  coord_polar(theta = "y", start=0)
```



Il ressort ainsi que les PME étudiées évoluent dans plusieurs filières notamment l'arachide, la mangue, l'anarcade et le riz à raison respectivement de 43.2 %, 23.2 %, 20.4 % et 13.2 % des PME.

Répartion des PME suivant le sexe du dirigeant, son niveau d'instruction, le statut juridique de l'entreprise et le statut de propriété face au local de travail

Pour faire cette analyse, nous utiliserons le package gtsummary pour n'avoir qu'un tableau.

```
## Répartion des PME suivant le sexe du dirigeant,
## son niveau d'instruction, le statut juridique de l'entreprise
## et le statut de propriété face au local de travail
library(gtsummary) # charger le package gtsummary

# faire le tableau
projet %>%
  dplyr::select(sexe, q24,q12, q25, q81,filiere) %>%
  gtsummary::tbl_summary(
    ## paramètres de tbl_summary
    by = sexe,
    ## variables qui forme les groupes: sexe
    label = list(
      sexe = "Sexe du dirigeant",
      q24 = "Age du dirigeant",
      q12 = "Statut juridique",
```

```
    q25 = "Niveau d'instruction du dirigeant",
    q81 = "Statut de propriété vis à vis du local de travail",
    filiere = "Filière"
  ),
  ## ajouter les étiquettes des variables
  percent = "column",
  ## Type de pourcentage affichés dans le tableau
  statistic = q24 ~ "{median}",
  ## statistiques à calculer pour l'âge du dirigeant (médiane)
  missing = "always",
  ## afficher les stat sur les valeurs manquantes
  missing_text = "Missing",
  ## formatage et nomination de la variable "valeur manquante"
) %>%
add_difference() %>%
## afficher la différence entre les groupes, le test de significativité de la différence
add_stat_label()
```

Characteristic	**Femme**, N = 191	**Homme**, N = 59	**Difference**	**95% CI**	**p-value**
Age du dirigeant, Median	56	50	4,065,644	-3,953,870, 12,085,158	0.3
Missing	0	0			
Statut juridique, n (%)			0.92	0.62, 1.2	
Association	3 (1.6%)	3 (5.1%)			
GIE	149 (78%)	30 (51%)			
Informel	32 (17%)	6 (10%)			
SA	1 (0.5%)	6 (10%)			
SARL	2 (1.0%)	11 (19%)			
SUARL	4 (2.1%)	3 (5.1%)			
Missing	0	0			
Niveau d'instruction du dirigeant, n (%)			0.88	0.58, 1.2	
Aucun niveau	70 (37%)	9 (15%)			
Niveau primaire	48 (25%)	8 (14%)			
Niveau secondaire	56 (29%)	18 (31%)			
Niveau Superieur	17 (8.9%)	24 (41%)			
Missing	0	0			
Statut de propriété vis à vis du local de travail, n (%)			0.17	-0.13, 0.46	
Locataire	16 (8.4%)	8 (14%)			
Propriétaire	175 (92%)	51 (86%)			
Missing	0	0			
Filière, n (%)			0.54	0.25, 0.84	
anacarde	34 (18%)	17 (29%)			
arachide	93 (49%)	15 (25%)			
mangue	38 (20%)	20 (34%)			
riz	26 (14%)	7 (12%)			
Missing	0	0			

```
  ## afficher une colonne qui signifie les statistiques calculées et leur format d'affichage. Ex: mean
```

Une analyse plus poussée des caractéristique des PME suivant le sexe de leur dirangeant révèle que la majeure partie des PME dirigées par les femmes sont des GIE (78 %) et des entreprises informelles (17 %) alors que chez leurs homologue masculins il s'agit plutôt de GIE (51 %)et de Société à responsabilités limitées (SARL 19%). Par ailleurs, les PME dirigées par les femmes évoluent plus dans les fillière arachide (49 %) et mangue (20 %) alors que celles dirigées par les hommes sont plus tournées vers les filières mangue (34 %) et anarcade (19 %).

Analyse du temps moyen mis pour recueillir les informations des PME

Pour analyser cela, nous allons d'abord créer une variable durre qui calcule le temps écoulé entre le debut de l'entretien et la soumission des information.

```
## Analyse du temps moyen mis pour recueillir les informations des PME
projet = projet %>% dplyr::mutate(duree = submissiondate - start)
### calcul de la durée minimale, maximale et moyenne
#### pour l'ensemble
tab = projet %>% summarise(min_duree = min(duree),
                     max_duree = max(duree),
                     mean_duree = mean(duree))

knitr::kable(tab, format = "markdown")
```

min_duree	max_duree	mean_duree
36.53333 mins	39828.53 mins	9463.662 mins

```
#### suivant le sexe
tableau_stat <- projet %>%
  group_by(sexe) %>%
  summarise(min_duree = min(duree),
            max_duree = max(duree),
            mean_duree = mean(duree))

knitr::kable(tableau_stat, format = "markdown")
```

sexe	min_duree	max_duree	mean_duree
Femme	36.53333 mins	39828.53 mins	8177.664 mins
Homme	119.25000 mins	39717.05 mins	13626.810 mins

Les résultats de la durée de soumission révèle qu'en moyenne les enquêteurs prennent en moyenne 9463.662 minutes soit 6 jours 13 heures pour soumettre les informations sur les PME. Par ailleurs, les information des PME dirigées par les hommes ont en moyenne une durée de soumission plus grande que celles sur les PME dirigées par les femmes.

Cartographie

Transformer du data.frame en données géographiques

```
# chargement des packages nécessaires
library(sf)
library(ggplot2)
library(ggspatial)
library(tmap)

# conversion du dataframe en objet géographique du format sf
projet_map <- projet  %>% sf::st_as_sf(coords= c("gps_menlongitude","gps_menlatitude"),crs = 4326)
```

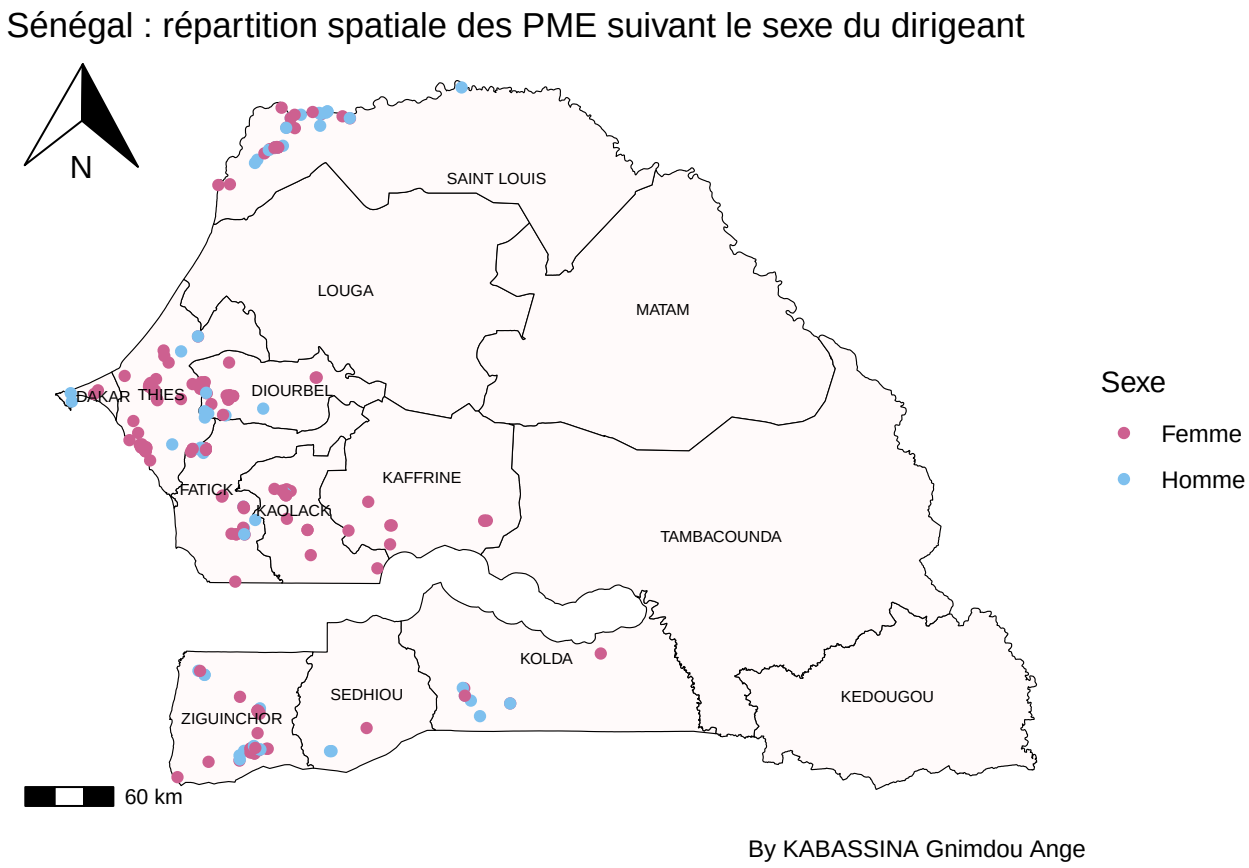
Répartition spaciale des PME suivant le sexe du dirigeant

Vu que les PME que nous souhaitons représenter proviennent d'Afrique de l'Ouest, nous allons d'abord représenter l'afrique de l'Ouest et y ajouter la position des PME par la suite.

```
## Répartition spaciale des PME suivant le sexe du dirigeant
### représentons d'abord le Sénégal
# Chargement des données sur le Sénégal
sen <- sf::st_read("donnees/Limite_Région.shp")
```

```
## Reading layer 'Limite_Région' from data source
##   'D:\ISEP3\S2\R\projet\KABASSINA_Ange_Projet R_ENSAE 2023\Partie_1\donnees\Limite_Région.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
```

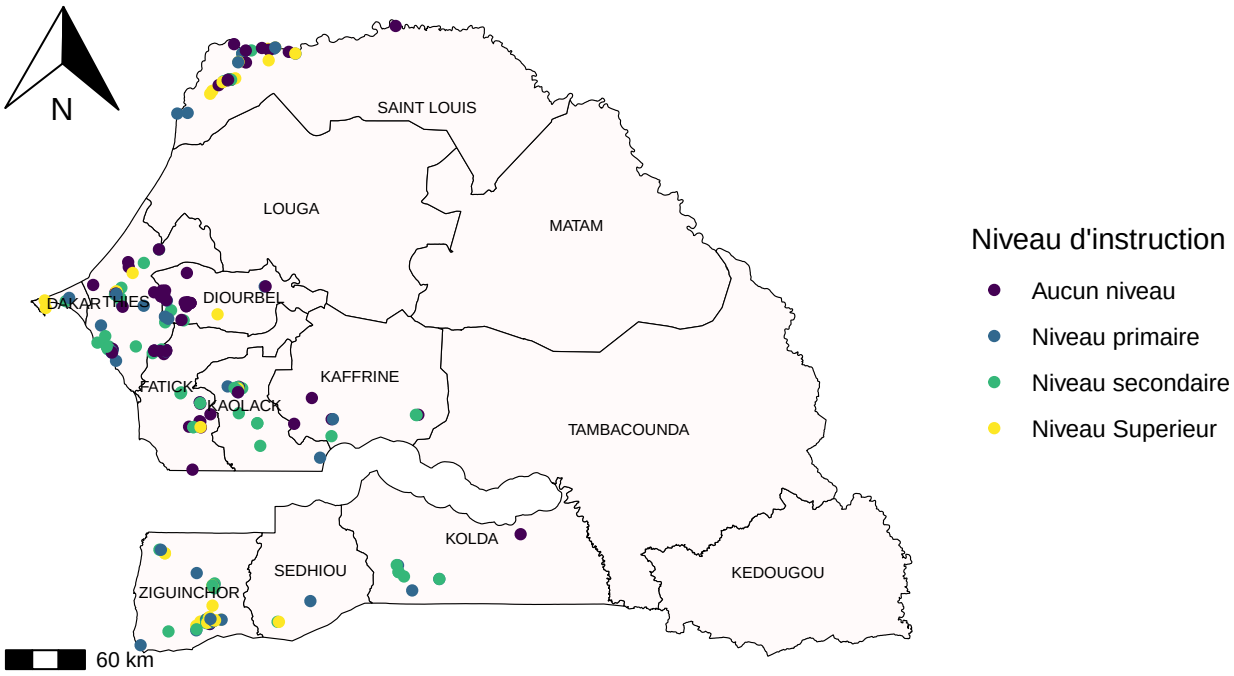
```
# Représentation graphique
ggplot() +
  geom_sf(data = sen, fill = "snow", color = "black") +
  geom_sf(data = projet_map, aes(color = sexe), size = 1.5) +
  scale_color_manual(values = c("hotpink3", "skyblue2")) +
  geom_sf_text(data = sen, aes(label = NOMREG), nudge_y = 0.2, size = 2) +
  annotation_scale(location = "bl", width_hint = 0.1) +
  annotation_north_arrow(location = "tl", which_north = "true") +
  theme_void() +
  theme(legend.position = "right") +
  labs(title = "Sénégal : répartition spatiale des PME suivant le sexe du dirigeant",
       caption = "By KABASSINA Gnimdou Ange",
       x = NULL, y = NULL,
       color = "Sexe")
```



Répartition spaciale des PME suivant le niveau d'instruction du dirigeant

```
## Répartition spaciale des PME suivant le niveau d'instruction du dirigeant
ggplot() +
  geom_sf(data = sen, fill = "snow", color = "black") +
  geom_sf(data = projet_map, aes(color = q25), size = 1.5) +
  scale_color_viridis_d() +
  geom_sf_text(data = sen, aes(label = NOMREG), nudge_y = 0.2, size = 2) +
  annotation_scale(location = "bl", width_hint = 0.1) +
  annotation_north_arrow(location = "tl", which_north = "true") +
  theme_void() +
  theme(legend.position = "right") +
  labs(title = "Sénégal : répartition spatiale des PME suivant le niveau d'instruction du dirigeant",
       caption = "By KABASSINA Gnimdou Ange",
       x = NULL, y = NULL,
       color = "Niveau d'instruction ")
```

Sénégal : répartition spatiale des PME suivant le niveau d'instruction du dirigeant

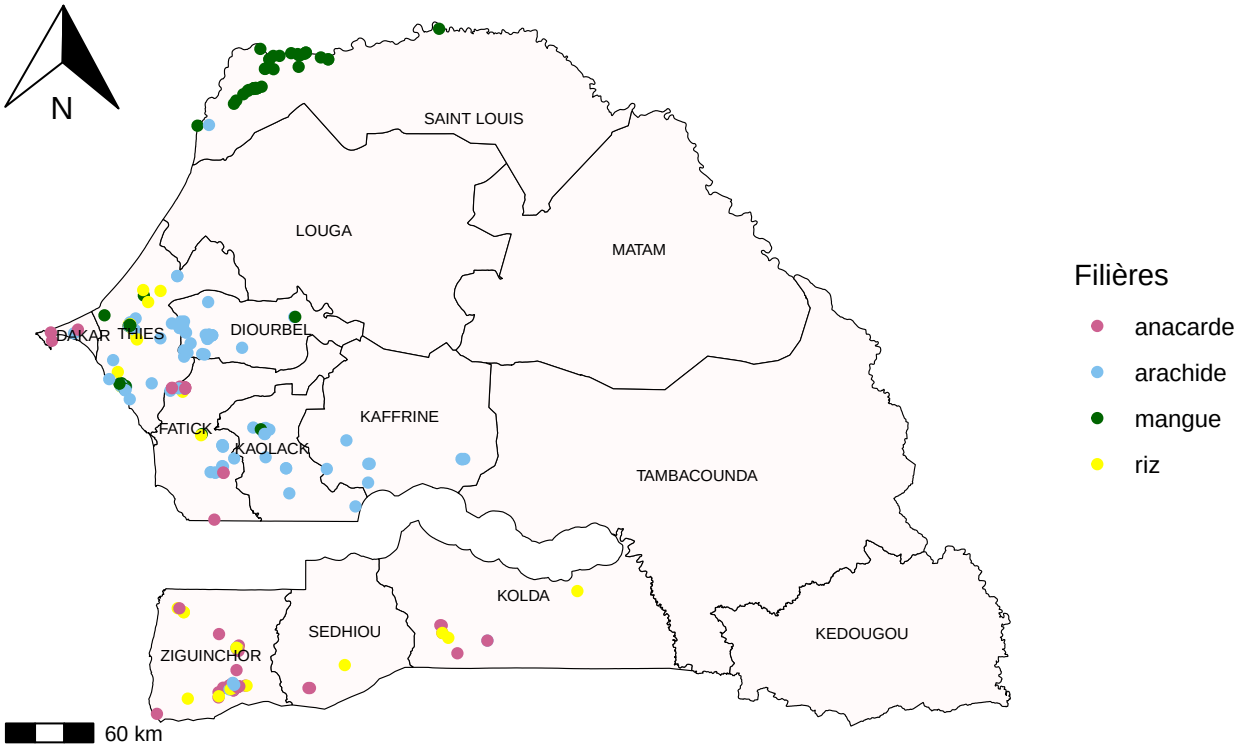


By KABASSINA Gnimdou Ange

Répartition spaciale des PME suivant la filière

```
# Répartition spaciale des PME suivant la filière
ggplot() +
  geom_sf(data = sen, fill = "snow", color = "black") +
  geom_sf(data = projet_map, aes(color = filiere), size = 1.5) +
  scale_color_manual(values = c("hotpink3", "skyblue2","darkgreen","yellow")) +
  geom_sf_text(data = sen, aes(label = NOMREG), nudge_y = 0.2, size = 2) +
  annotation_scale(location = "bl", width_hint = 0.1) +
  annotation_north_arrow(location = "tl", which_north = "true") +
  theme_void() +
  theme(legend.position = "right") +
  labs(title = "Sénégal : répartition spatiale des PME suivant les filières d'activité",
       caption = "By KABASSINA Gnimdou Ange",
       x = NULL, y = NULL,
       color = "Filières")
```

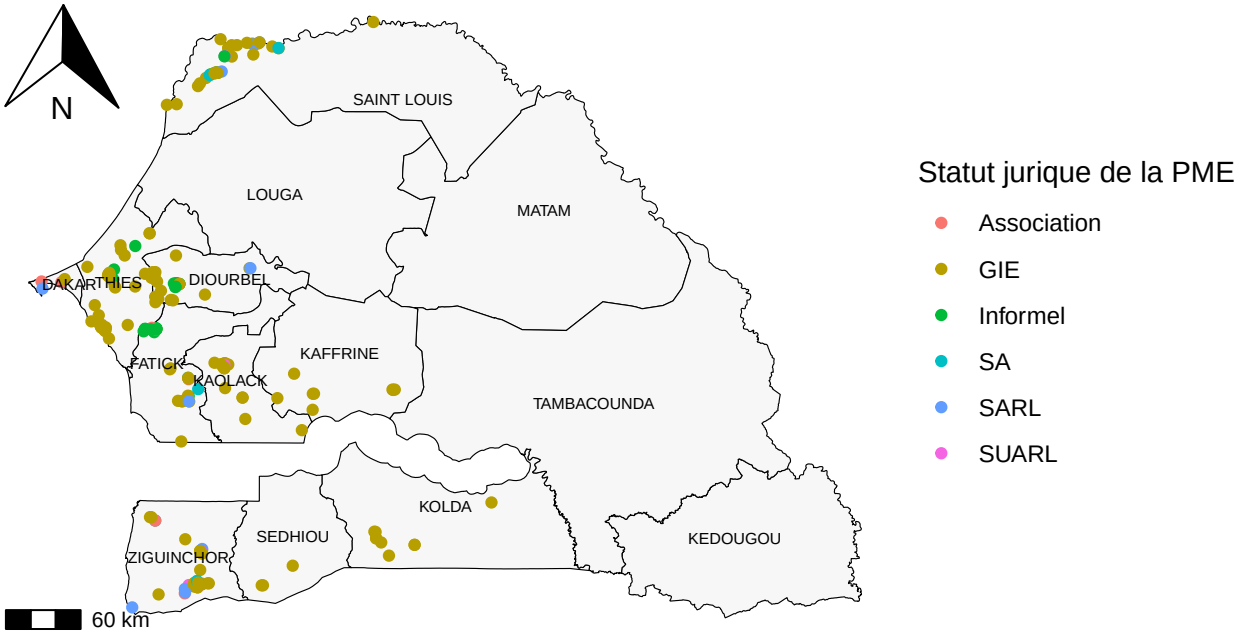
Sénégal : répartition spatiale des PME suivant les filières d'activité



Répartition spaciale des PME suivant le statut jurique de la PME

```
# Répartition spaciale des PME suivant le statut jurique de la PME
ggplot() +
  geom_sf(data = sen, fill = "gray97", color = "black") +
  geom_sf(data = projet_map, aes(color = q12), size = 1.5) +
  scale_fill_manual(values = cm.colors(6))+
  geom_sf_text(data = sen, aes(label = NOMREG), nudge_y = 0.2, size = 2) +
  annotation_scale(location = "bl", width_hint = 0.1) +
  annotation_north_arrow(location = "tl", which_north = "true") +
  theme_void() +
  theme(legend.position = "right") +
  labs(title = "Sénégal : répartition spatiale des PME suivant le statut jurique de la PME",
       caption = "By KABASSINA Gnimdou Ange",
       x = NULL, y = NULL,
       color = "Statut jurique de la PME")
```

Sénégal : répartition spatiale des PME suivant le statut jurique de la PME



By KABASSINA Gnimdou Ange