

TP1-Logiciel statistique R

Traoré Sié Rachid

2024-04-02

Contents

1	IMPORTATION ET MISE EN FORME	2
1.1	Importons la base de données dans un objet de type data.frame nommé projet	2
1.2	Donnons le nombre de lignes (i.e. le nombre de PME) et colonnes (i.e. nombre de variables) de la base projet	2
1.3	Vérifions s'il y a des valeurs manquantes pour la variable key dans la base projet. Si oui, identifier la (ou les) PME concernée(s).	2
2	CREATION DES VARIABLES	3
2.1	Rénommons la variable q1 en region	3
2.2	Rénommons la variable q2 en departement	3
2.3	Rénommons la variable q23 en sexe	3
2.4	Créons la variable sexe_2 qui vaut 1 si sexe égale à Femme et 0 sinon.	4
2.5	Créons un data.frame nommé langues qui prend les variables key et les variables correspondantes décrites plus haut.	4
2.6	Créons une variable parle qui est égale au nombre de langue parlée par le dirigeant de la PME.	4
2.7	Sélectionnez uniquement les variables key et parle, l'objet de retour sera langues	4
2.8	Merger les data.frame projet et langues	4
3	ANALYSE DESCRIPTIVE	5
3.1	la répartition des PME suivant:	5
3.2	Faisons les statistiques descriptives de notre choix sur les autres variables	7

```
#-----  
library(readxl) # pour importer la base Base_Projet.xlsx  
#-----  
library(dplyr) # pour manipuler la base Base_Projet.xlsx  
#-----  
library(janitor) # pour réaliser les tableaux de contingence avec la fonction tabyl
```

1 IMPORTATION ET MISE EN FORME

1.1 Importons la base de données dans un objet de type data.frame nommé projet

```
#-----  
projet <- read_excel("Base_Projet.xlsx",  
  sheet = "Feuil1") # importation de la base de la feuille Feuil1  
#-----  
View(projet) # visualier la base
```

1.2 Donnons le nombre de lignes (i.e. le nombre de PME) et colonnes (i.e. nombre de variables) de la base projet

- le nombre de lignes (i.e. le nombre de PME) de la base projet

```
#-----  
nrow(projet)
```

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

```
#-----  
# nrow permet de renvoyer le nombre de lignes de la base de données
```

- le nombre de colonnes (i.e. nombre de variables) de la base projet

```
#-----  
ncol(projet)
```

```
## [1] 33
```

```
#-----  
# ncol permet de renvoyer le nombre de colonnes de la base de données
```

1.3 Vérifions s'il y a des valeurs manquantes pour la variable key dans la base projet. Si oui, identifier la (ou les) PME concernée(s).

```
# Vérification des valeurs manquantes dans la variable key  
valeurs_manquantes <- is.na(projet$key)  
  
# Affichage des indices des valeurs manquantes  
indices_manquants <- which(valeurs_manquantes)  
print(indices_manquants)
```

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

Il n'y a pas donc pas de valeurs manquantes pour la variable key dans la base projet.

2 CREATION DES VARIABLES

2.1 Rénommons la variable q1 en region

```
#-----  
projet <- projet %>%  
  rename(region=q1)  
#-----  
#rename() permet de renommer les colonnes de nos bases de données  
#-----  
#rename(nouvelle_variable=ancienne_variable)  
#-----  
#nouvelle_variable est region  
#-----  
#ancienne_variable est q1  
#-----
```

2.2 Rénommons la variable q2 en departement

```
#-----  
projet <- projet %>%  
  rename(departement=q2)  
#-----  
#rename() permet de renommer les colonnes de nos bases de données  
#-----  
#rename(nouvelle_variable=ancienne_variable)  
#-----  
#nouvelle_variable est departement  
#-----  
#ancienne_variable est q2
```

2.3 Rénommons la variable q23 en sexe

```
#-----  
projet <- projet %>%  
  rename(sexe=q23)  
#-----  
#rename() permet de renommer les colonnes de nos bases de données  
#-----  
#rename(nouvelle_variable=ancienne_variable)  
#-----  
#nouvelle_variable est sexe  
#-----  
#ancienne_variable est q23  
#-----
```

2.4 Créons la variable `sexe_2` qui vaut 1 si `sexe` égale à `Femme` et 0 sinon.

```
#-----
projet <- projet %>%
  mutate(sexe_2=ifelse(sexe=="Femme",1,0))
#-----
# mutate crée une autre variable sexe_2
# où l'on code la modalité Femme en 1 et en 0 sinon.
```

2.5 Créons un `data.frame` nommé `langues` qui prend les variables `key` et les variables correspondantes décrites plus haut.

```
#-----
langues <- projet %>%
  select(key,starts_with("q24a_"))
#-----
View(langues)
#-----
# select permet ici de sélectionner la variable key
# et toutes variables commençant par q24a_ avec startswith
```

2.6 Créons une variable `parle` qui est égale au nombre de langue parlée par le dirigeant de la PME.

```
#-----
langues$parle <- langues[,2:10] %>%
  apply(MARGIN = 1,FUN = sum)
#-----
# Il s'agit de créer une variable parle
# les variables concernant les langues sont à la position de 2 à 10
# apply permet de sommer avec FUN=sum suivant la ligne grâce à MARGIN=1
```

2.7 Sélectionnez uniquement les variables `key` et `parle`, l'objet de retour sera `langues`

```
#-----
langues <- langues %>%
  select(key,parle)
#-----
# select permet ici de sélectionner la variable key et la variable parle
```

2.8 Merger les `data.frame` `projet` et `langues`

```
#-----
merged_projet <- full_join(projet,langues,by="key")
#-----
View(merged_projet)
#-----
# full_join permet d'effectuer la jointure des deux
# data_frames projet et langues selon la variable key.
```

3 ANALYSE DESCRIPTIVE

3.1 la répartition des PME suivant:

- Selon le sexe

```
#-----
merged_projet %>%
  tabyl(sexe) %>%
  adorn_pct_formatting(digits = 2)
```

```
##   sexe    n percent
##  Femme 191  76.40%
##  Homme  59  23.60%
```

```
#-----
# tabyl permet de fournir la répartition des PME selon le sexe des responsables .
```

Parmi les représentants de chaque PME, on a 191 femmes et 59 hommes.

- Selon le niveau d’instruction

```
#-----
merged_projet %>%
  tabyl(q25) %>%
  adorn_pct_formatting(digits = 2)
```

```
##               q25    n percent
##   Aucun niveau 79   31.60%
## Niveau Supérieur 41   16.40%
## Niveau primaire 56   22.40%
## Niveau secondaire 74  29.60%
```

```
#-----
# tabyl permet de fournir la répartition des PME
# selon le niveau d'instruction des responsables représenté par la variable q25.
```

Parmi les représentations des PME, 79 n’ont aucun niveau, 56 ont un niveau primaire, 74 ont un niveau secondaire et 41 ont un niveau supérieur.

- Selon le statut juridique

```
#-----
merged_projet %>%
  tabyl(q12) %>%
  adorn_pct_formatting(digits = 2)
```

```
##      q12    n percent
## Association    6   2.40%
##      GIE  179  71.60%
## Informel    38  15.20%
##      SA     7   2.80%
##      SARL   13   5.20%
##      SUARL    7   2.80%
```

```
#-----
# tabyl permet de fournir la répartition des PME
# selon le statut juridique représenté par q12.
```

Parmi les représentations des PME, 6 sont des associations, 179 sont des GIE , 38 sont informels , 7 sont des SA , 13 sont des SARL et 7 sont SUARL.

- Selon le propriétaire/locataire

```
#-----
merged_projet %>%
  tabyl(q81) %>%
  adorn_pct_formatting(digits = 2)
```

```
##      q81    n percent
## Locataire   24   9.60%
## Propriétaire 226  90.40%
```

```
#-----
# tabyl permet de fournir la répartition des PME
# selon propriétaire/locataire représenté par q81.
```

Parmi les représentants des PME, 24 sont locataires 226 sont propriétaires

- Selon le statut juridique et le sexe

```
#-----
merged_projet %>%
  tabyl(q12,sexe)
```

```
##      q12 Femme Homme
## Association    3     3
##      GIE   149    30
## Informel     32     6
##      SA      1     6
##      SARL     2    11
##      SUARL    4     3
```

```
#-----
# tabyl permet de fournir la répartition des PME selon le statut juridique et le sexe
# des responsables représenté par la q12 et sexe.
```

- Selon le niveau d'instruction et le sexe

```
#-----
merged_projet %>%
  tabyl(q25,sexe)
```

```
##           q25 Femme Homme
##   Aucun niveau    70     9
##   Niveau primaire  48     8
##   Niveau secondaire 56    18
##   Niveau Supérieur 17    24
```

```
#-----
# tabyl permet de fournir la répartition des PME selon le niveau d'instruction et le sexe
# des responsables représenté par q25 et sexe.
```

- Selon Propriétaire/locataire suivant le sexe

```
#-----
merged_projet %>%
  tabyl(q81,sexe)
```

```
##           q81 Femme Homme
##   Locataire    16     8
##   Propriétaire 175    51
```

```
#-----
# tabyl permet de fournir la répartition des PME selon le Propriétaire/locataire et le sexe
# des responsables représenté par q81 et sexe.
```

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

- Analyse de la filière arachide

Le nombre de PME dans la filière arachide est:

```
#-----
merged_projet %>%
  select(filiere_1) %>%
  apply(MARGIN = 2,FUN=sum)
```

```
## filiere_1
##         108
```

```
#-----
#select selectionne la variable filiere_2 representant la filiere anacarde
#-----
# apply permet de sommer avec FUN=sum les valeurs de la colonne filiere_1 grâce à MARGIN=1 .
```

- Analyse de la filière anacarde

Le nombre de PME dans la filière anacarde est:

```
#-----
merged_projet %>%
  select(filiere_2) %>%
  apply(MARGIN = 2,FUN=sum)
```

```
## filiere_2
##          61
```

```
#-----
#select selectionne la variable filiere_2 representant la filiere anacarde
#-----
# apply permet de sommer avec FUN=sum les valeurs de la colonne filiere_2 grâce à MARGIN=1 .
```

- Analyse de la filière mangue

Le nombre de PME dans la filière mangue est:

```
#-----
merged_projet %>%
  select(filiere_3) %>%
  apply(MARGIN = 2,FUN=sum)
```

```
## filiere_3
##          89
```

```
#-----
# apply permet de sommer avec FUN=sum les valeurs de la colonne filiere_3 grâce à MARGIN=1 .
```

- Analyse de la filière riz

Le nombre de PME dans la filière riz est:

```
#-----
merged_projet %>%
  select(filiere_4) %>%
  apply(MARGIN = 2,FUN=sum)
```

```
## filiere_4
##          92
```



```
#-----
# apply permet de sommer avec FUN=sum les valeurs de la colonne filiere_4 grâce à MARGIN=1 .
```

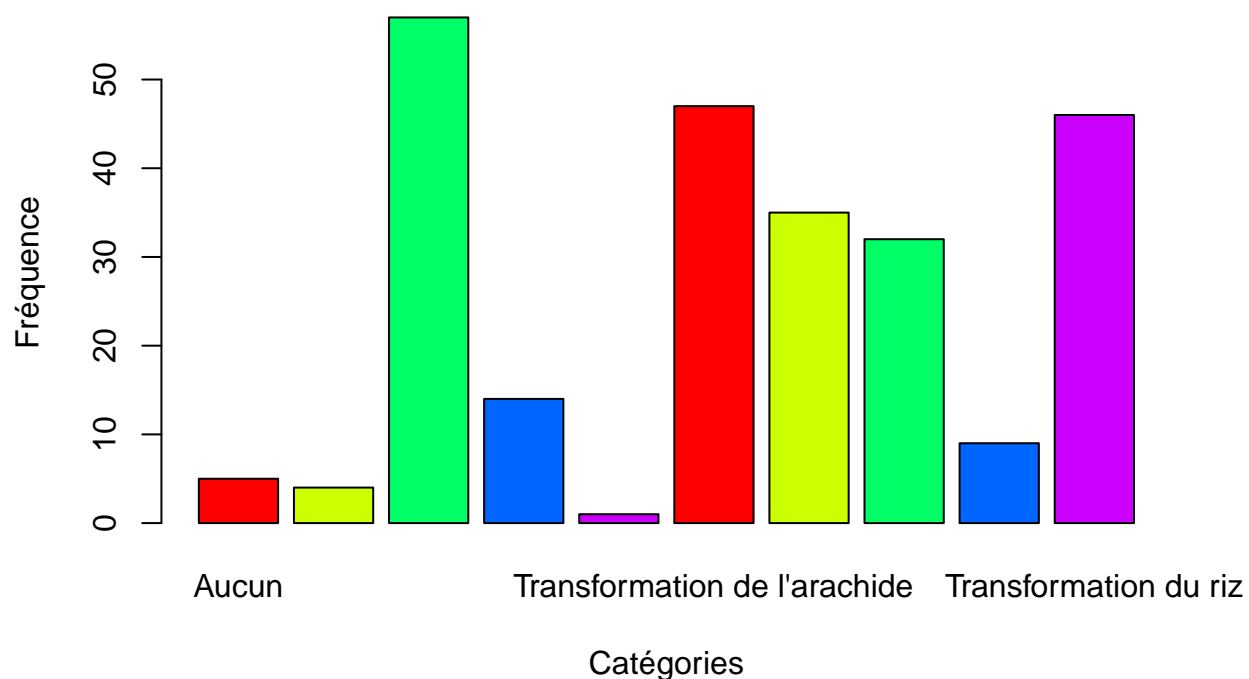
- Analyse des activités principales des entreprises

```
#-----
merged_projet %>%
  tabyl(q8) %>%
  adorn_pct_formatting(digits = 2) # répartition des entreprises selon les activités principales
```

```
##              q8  n percent
##              Aucun  5   2.00%
##              Autre a preciser  4   1.60%
##      Transformation d'autres céréales  57  22.80%
##      Transformation d'autres fruits et legumes  14   5.60%
##      Transformation d'autres produits oléagineux  1   0.40%
##              Transformation de l'arachide  47  18.80%
##              Transformation de la mangue  35  14.00%
##      Transformation de la noix de cajoux  32  12.80%
##      Transformation de la pomme de cajoux  9   3.60%
##              Transformation du riz  46  18.40%
```

```
barplot(table(merged_projet$q8),
  main = "Repartition des activités principales des entreprises", # le titre du graphique
  xlab = "Catégories", ylab = "Fréquence", # Les titres des axes
  col = rainbow(5)) # pour la couleur
```

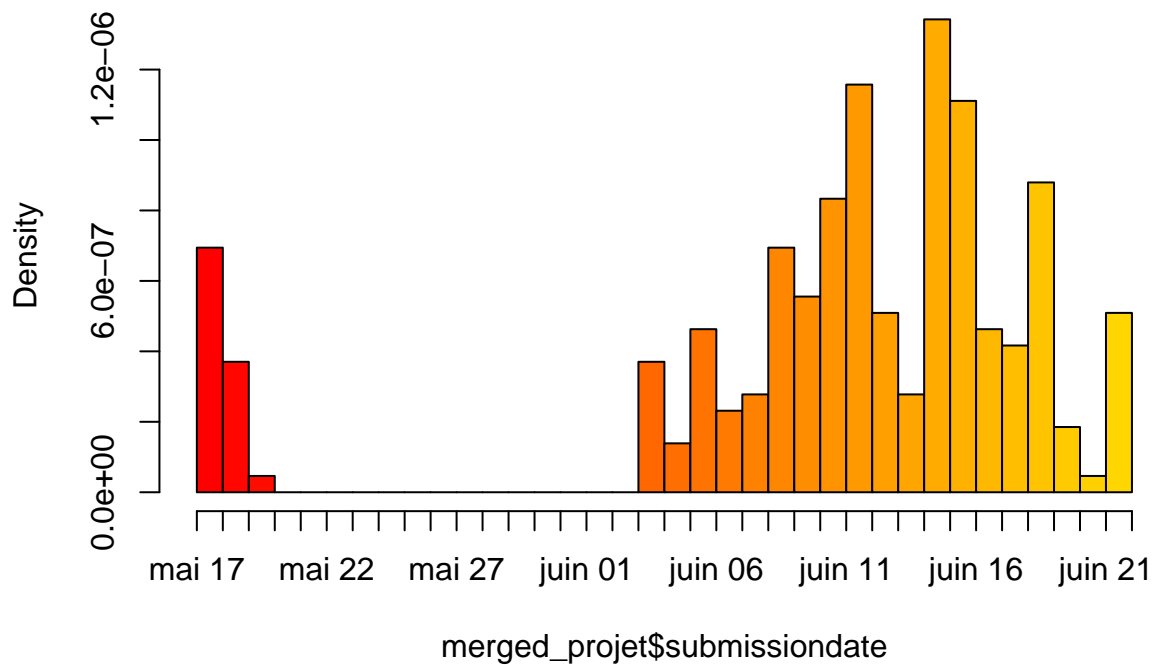
Repartition des activités principales des entreprises



- Analyse des dates de soumission

```
hist(merged_projet$submissiondate, # la variable
     breaks = "day", # Le mode de répartition
     main = "Distribution temporelle des données", # le titre du graphique
     col=rainbow(250)) # pour la couleur
```

Distribution temporelle des données



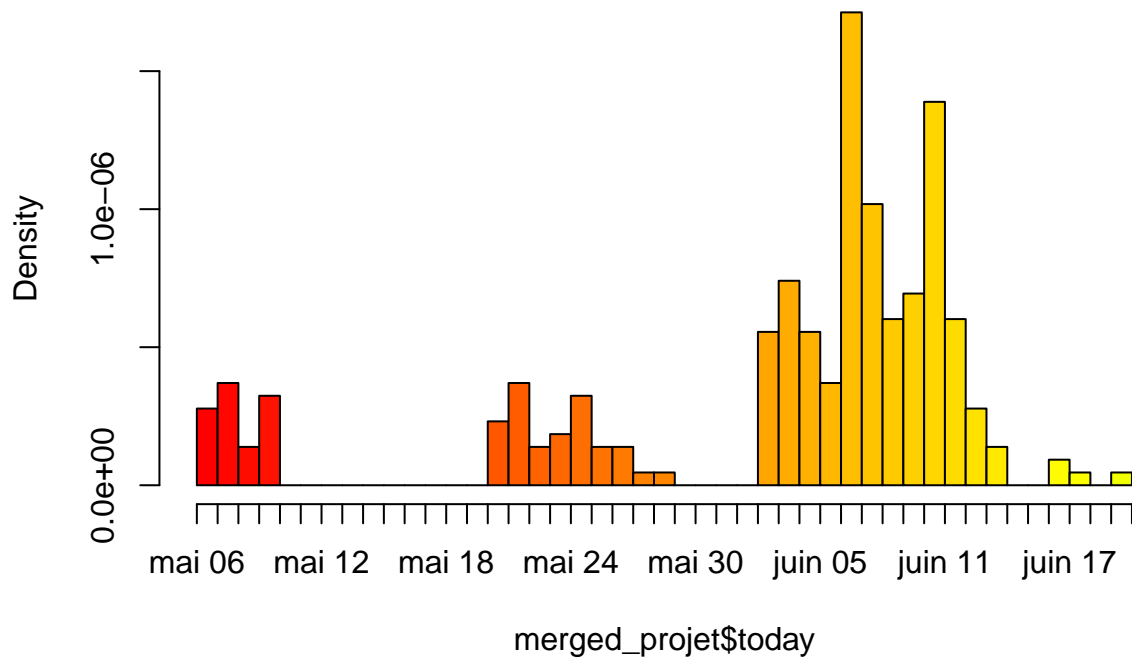
```
# Histogramme des dates
```

L'analyse de la distribution temporelle des dates de soumission des informations montre que les soumissions ont été effectuées du 17 mai au 20 mai puis du 04 juin au 22 juin. Entre le 20 mai et le 04 juin, il n'a pas eu de soumissions.

- **Analyse des dates d'enquête

```
hist(merged_projet$today,  
     breaks = "day",  
     main = "Distribution temporelle des données",  
     col=rainbow(250))
```

Distribution temporelle des données



```
# Histogramme des dates
```

Concernant les dates d'enquêtes, les jours correspondant sont du 06 mai au 10 mai puis du 22 mai au 30 mai ,du 04 juin au 16 juin et 18 juin au 21 juin

- Les statistiques descriptives univariées

```
# Fonction pour les statistiques descriptives univariées avec graphiques
```

```
univarie <- function(data,variable,plot = TRUE) {
```

```
  statistics <- data %>%
```

```
    tabyl(variable) # pour la repartition de la variable
```

```
  nb <- length(unique(data[[variable]]))
```

```
  if (plot) {
```

```
    barplot(table(data[[variable]]),
```

```
      main = "Histogramme des données",
```

```
      xlab = "Valeurs", col = rainbow(nb))
```

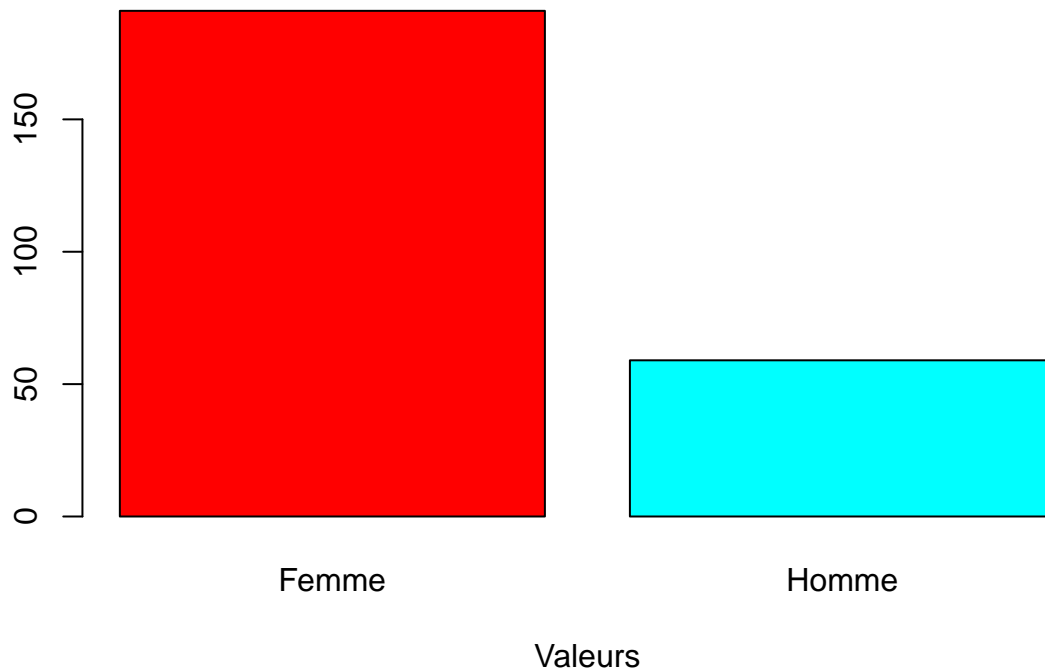
```
  }
```

```
  return(statistics)
```

```
}
```

```
sexe <- merged_projet$sexe
univarie(merged_projet,"sexe",plot=T)
```

Histogramme des données



```
##  sexe  n percent
##  Femme 191  0.764
##  Homme  59  0.236
```

```
# Fonction pour les statistiques descriptives bivariées avec graphiques

bivarie <- function(data,variable1,variable2, plot = TRUE) {

  # tableau de contingence

  tableau_croise <- table(data[[variable1]],data[[variable2]])

  if (plot) {

    # Graphique en barres empilées

    barplot(tableau_croise,
             main = "Distribution",
             xlab = variable2, ylab = "Fréquence",
             col = c("blue", "red"))
    legend("topright", legend = rownames(tableau_croise), fill = c("blue", "red"))
  }
}
```

```

}

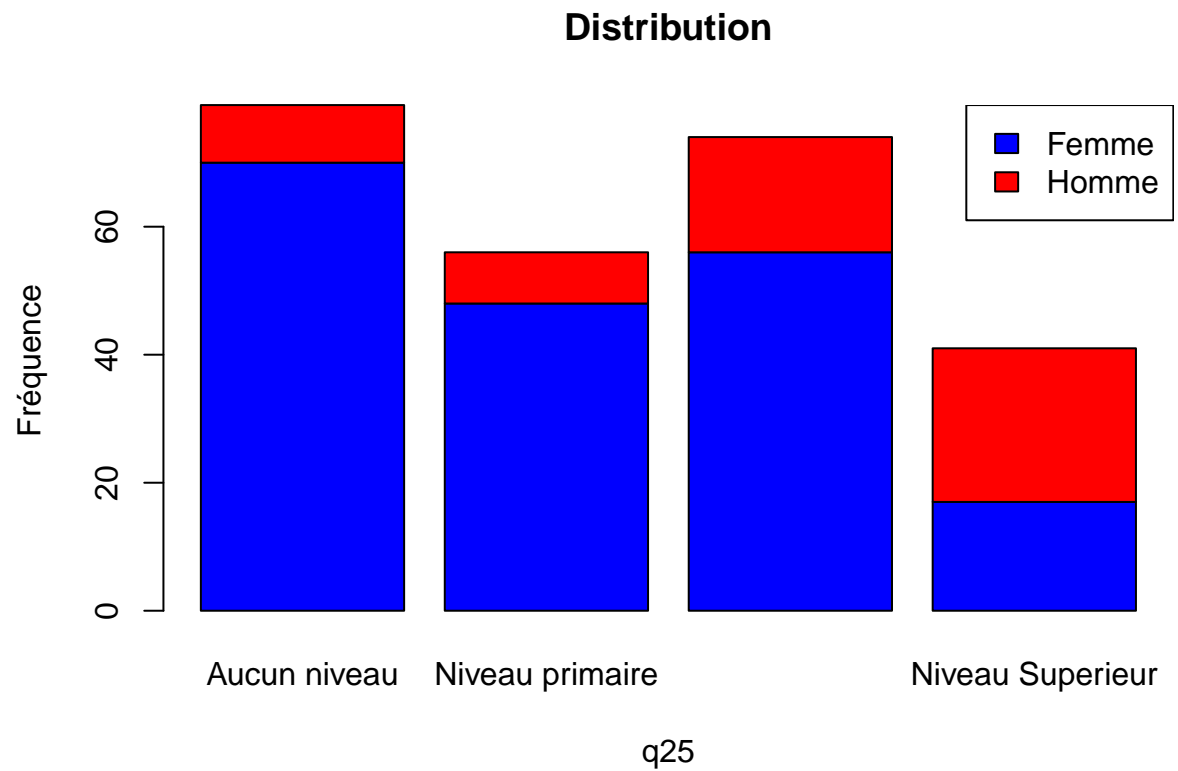
return(tableau_croise)
}

```

```

bivarie(merged_projet,"sexe","q25")

```



```

##
##      Aucun niveau Niveau primaire Niveau secondaire Niveau Supérieur
##  Femme          70          48          56          17
##  Homme           9           8          18          24

```