Rapport

DIOP_Ousseynou

2023-05-07

Question 1-2-3-4

```
knitr::opts_chunk$set(echo=TRUE, error=FALSE, warning=FALSE, message=TRUE)
CA<- "C:\\Users\\dell\\Desktop\\ENSAE\\ISEP2\\Semestre_2\\Programmation R\\Devoir\\Traitement"
df_Diop<- read.csv2(paste0(CA,"\\df_Diop.csv"))
head(df_Diop)
```

```
##   age      moy salaire nombre_pas jour_absent      nom genre
## 1  22 17.09545   99130        840          12 Floriane Homme
## 2  41 16.60973  147254       1370           4    Chloé Femme
## 3  29 14.14334   86367       1314           9   Hélène Femme
## 4  42 15.35980   94743       1321           3   Pierre Homme
## 5  22 18.86301   74952       1916          16      Lou Femme
## 6  21 16.50604  111850       1312           9   Benoit Homme
##          niveau_etude lettre_preferee voyage_etude
## 1 diplôme d'ingénieur               O          non
## 2           doctorat               E          non
## 3                bac               G          non
## 4              bac+3               G          oui
## 5                CAP               F          non
## 6                CAP               A          non
```

```
tail(df_Diop)
```

```
##    age      moy salaire nombre_pas jour_absent     nom genre
## 94  23 15.38383  119071       1746           9 Nolwenn Femme
## 95  33 14.91305  181854       1153           2 Mathieu Femme
## 96  19 14.88698   78926       1986          24  Audrey Homme
## 97  40 17.01717   88122       1571           5  Audrey Femme
## 98  43 15.17188  107176       1202          18 Pauline Femme
## 99  18 17.57527  181972       1269          14 Olivier Homme
##                   niveau_etude lettre_preferee voyage_etude
## 94 diplôme d'école de commerce               F          oui
## 95                      bac+2               F          non
## 96                        CAP               M          non
## 97                        CAP               F          oui
## 98                      bac+5               D          oui
## 99       master professionnel               M          oui
```

```
View(df_Diop)
```

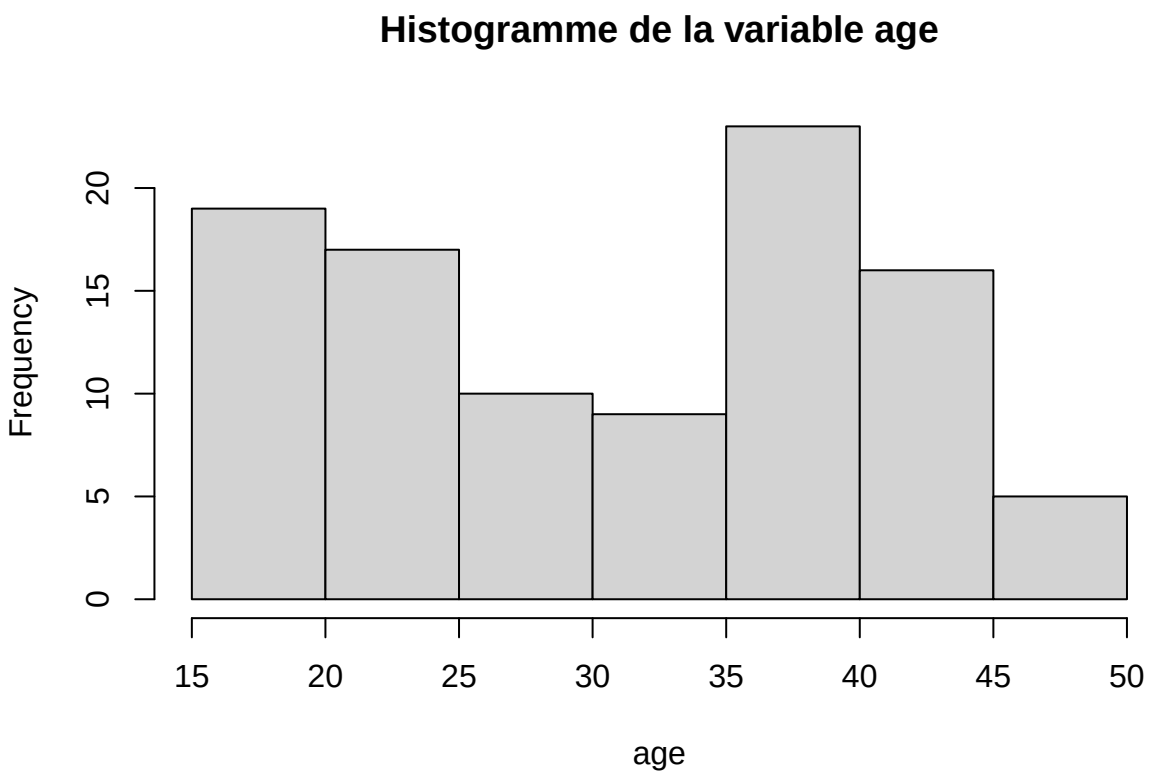
Question 5

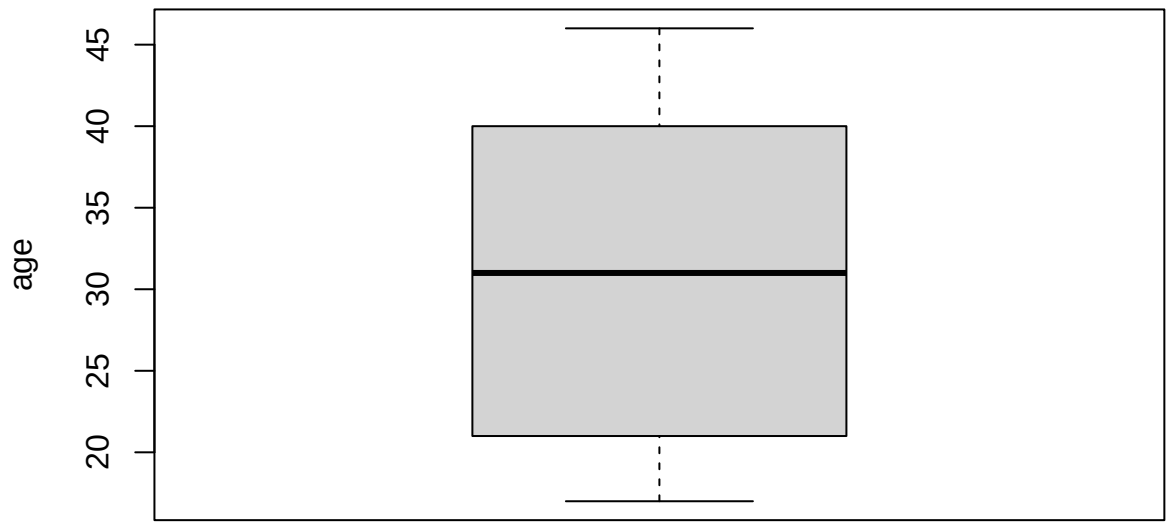
```
knitr::opts_chunk$set(echo=TRUE, error=FALSE, warning=FALSE, message=TRUE)
d.var.quant <- function(baseD, var){
  library(ggplot2)
  # Calcul des tendances centrales
  moy <- mean(baseD[[var]])
  med <- median(baseD[[var]])
  et <- sd(baseD[[var]])

  # Affichage des tendances centrales
  cat("Tendances centrales:\n")
  cat(paste0("Moyenne: ", round(moy, 2), "\n"))
  cat(paste0("Médiane: ", med, "\n"))
  cat(paste0("Écart-type: ", round(et, 2), "\n"))

  # Graphiques
  hist(baseD[[var]], main = paste0("Histogramme de la variable ", var), xlab = var)
  boxplot(baseD[[var]], main = paste0("Boxplot de la variable ", var), ylab = var)

  # Intervalle de confiance
  conf_int <- t.test(baseD[[var]], conf.level = 0.95)$conf.int
  cat(paste0("Intervalle de confiance à 95%: [", round(conf_int[1], 2), ", ",   round(conf_int[2], 2),

  print(hist)
  print(boxplot)

}
```

age

```
d.var.quant(df_Diop,"age")
```

```
## Tendances centrales:
## Moyenne: 31.24
## Médiane: 31
## Écart-type: 9.58
```


Histogramme de la variable age
Frequency
0
5
10
15
20
15
20
25
30
35
40
45
50
age

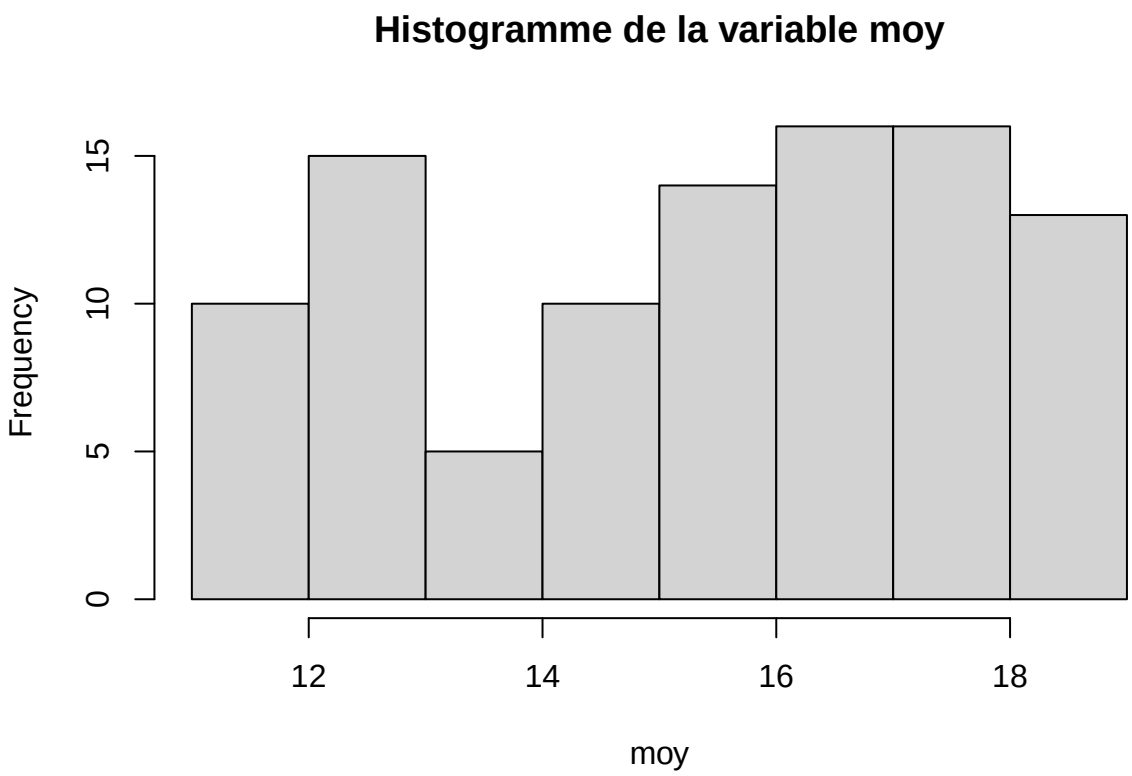


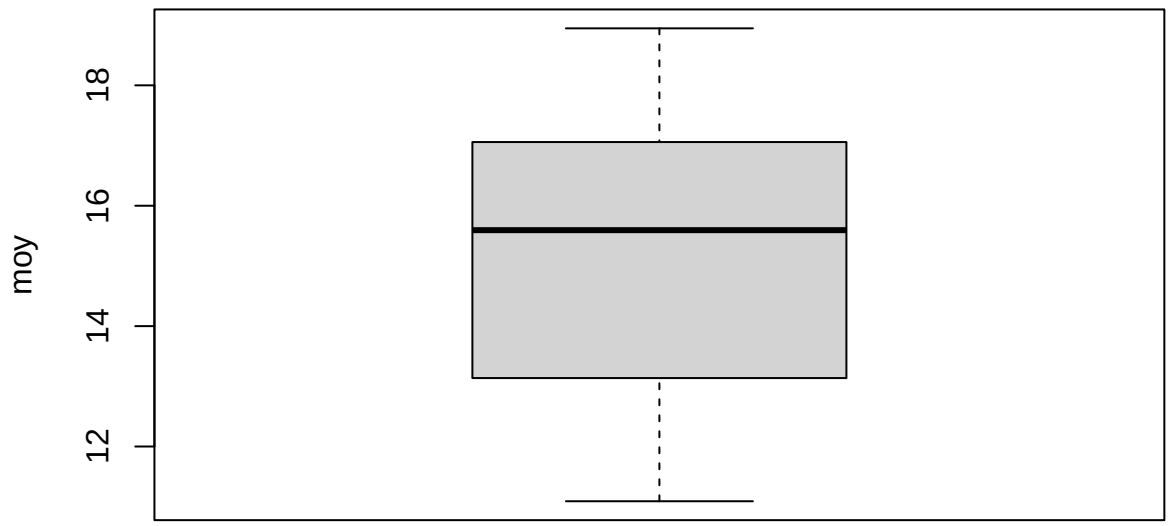

```
## Intervalle de confiance à 95%: [29.33, 33.15]
## function (x, ...)
## UseMethod("hist")
## <bytecode: 0x0000019a319062b8>
## <environment: namespace:graphics>
## function (x, ...)
## UseMethod("boxplot")
## <bytecode: 0x0000019a3194f068>
## <environment: namespace:graphics>
```

moy

```
d.var.quant(df_Diop,"moy")
```

```
## Tendances centrales:
## Moyenne: 15.33
## Médiane: 15.5939283370972
## Écart-type: 2.3
```


Histogramme de la variable moy
Frequency
0
5
10
15
12
14
16
18
moy

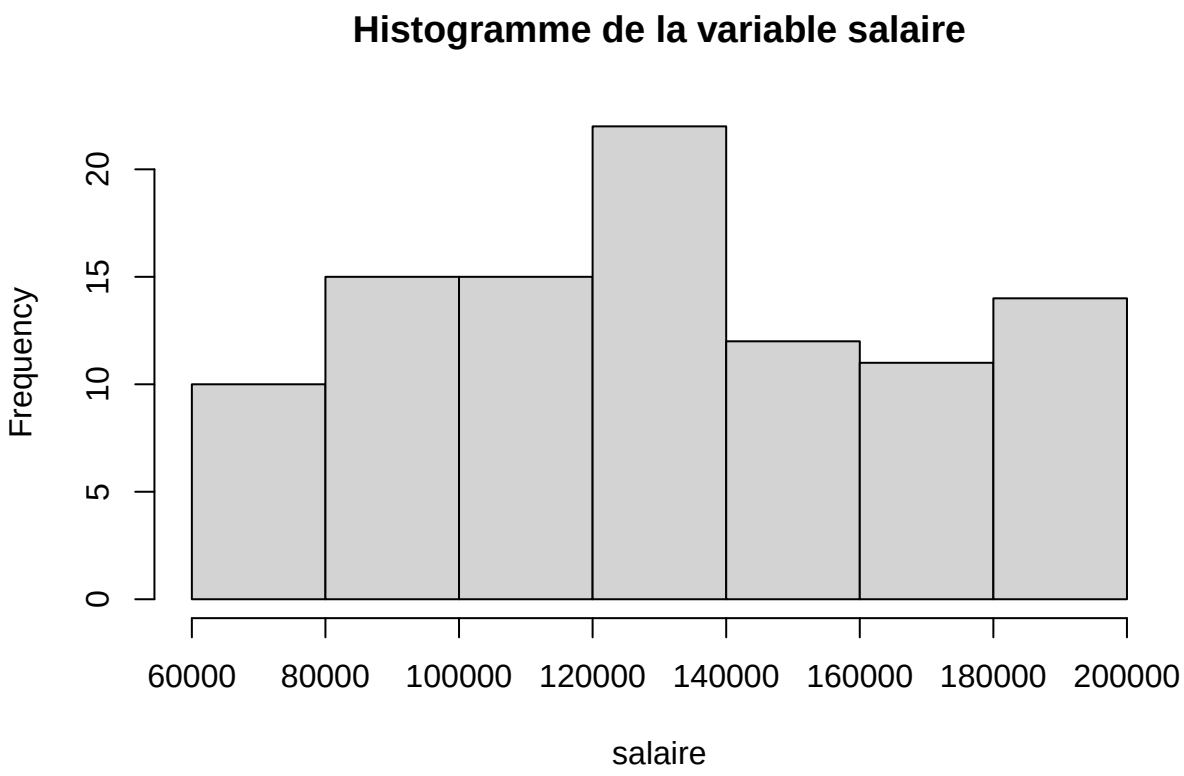


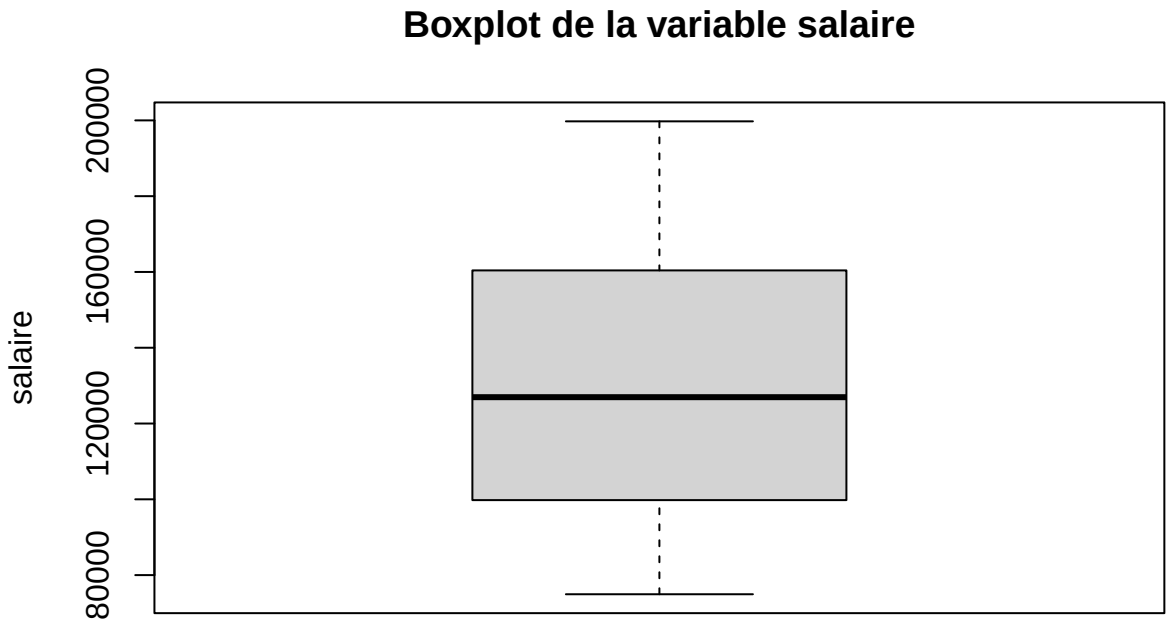

```
## Intervalle de confiance à 95%: [14.87, 15.79]
## function (x, ...)
## UseMethod("hist")
## <bytecode: 0x0000019a319062b8>
## <environment: namespace:graphics>
## function (x, ...)
## UseMethod("boxplot")
## <bytecode: 0x0000019a3194f068>
## <environment: namespace:graphics>
```

salaire

```
d.var.quant(df_Diop,"salaire")
```

```
## Tendances centrales:
## Moyenne: 130739.49
## Médiane: 126951
## Écart-type: 36989.52
```


Histogramme de la variable salaire
Frequency
0
5
10
15
20
60000
80000
100000
120000
140000
160000
180000
200000
salaire

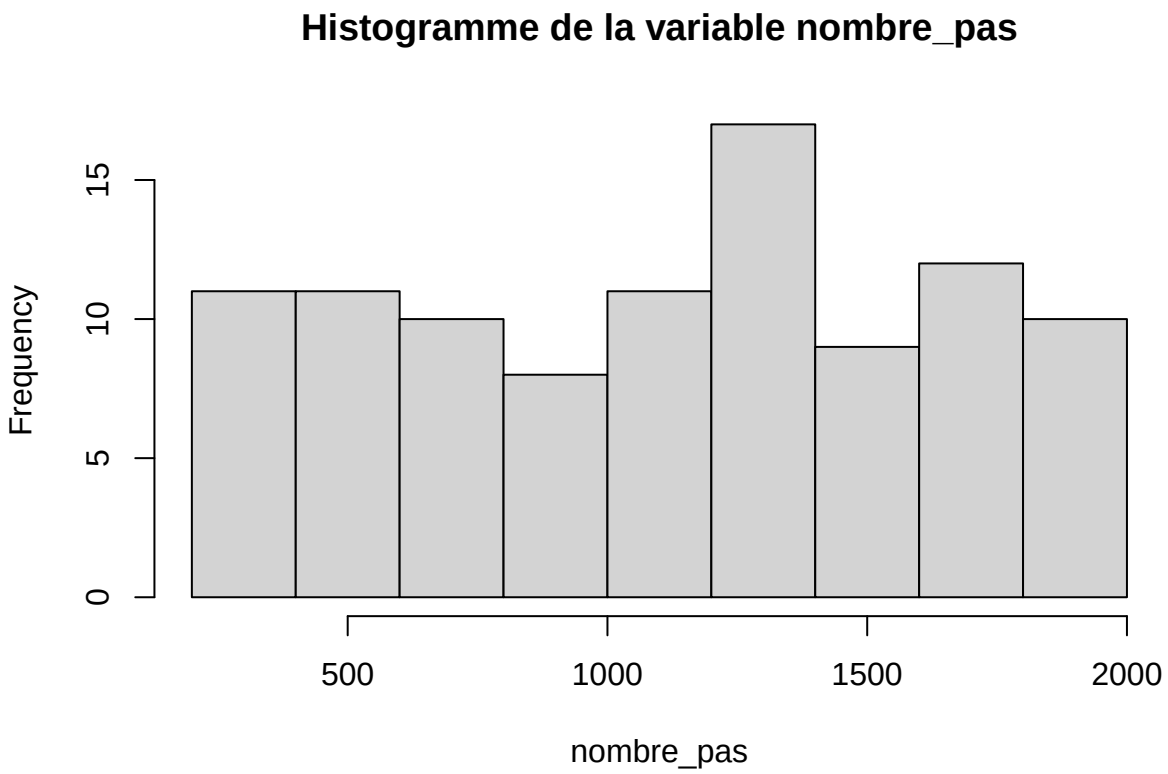


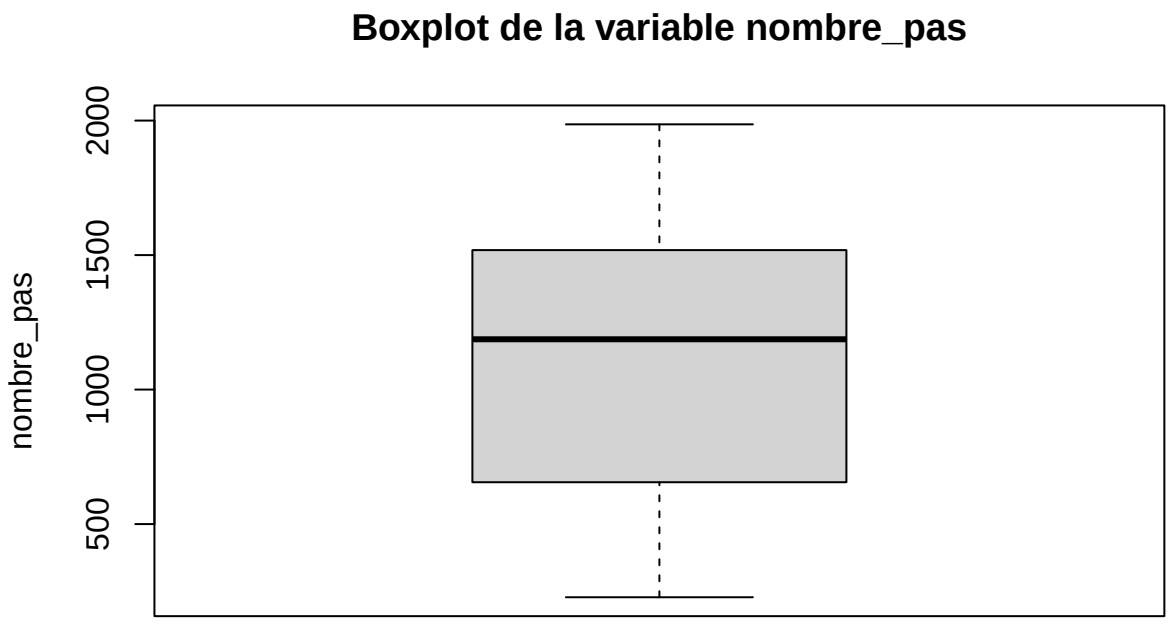

```
## Intervalle de confiance à 95%: [123362.06, 138116.93]
## function (x, ...)
## UseMethod("hist")
## <bytecode: 0x0000019a319062b8>
## <environment: namespace:graphics>
## function (x, ...)
## UseMethod("boxplot")
## <bytecode: 0x0000019a3194f068>
## <environment: namespace:graphics>
```

nombre_pas

```
d.var.quant(df_Diop,"nombre_pas")
```

```
## Tendances centrales:
## Moyenne: 1114.02
## Médiane: 1187
## Écart-type: 513.87
```


Histogramme de la variable nombre_pas
Frequency
0
5
10
15
500
1000
1500
2000
nombre_pas

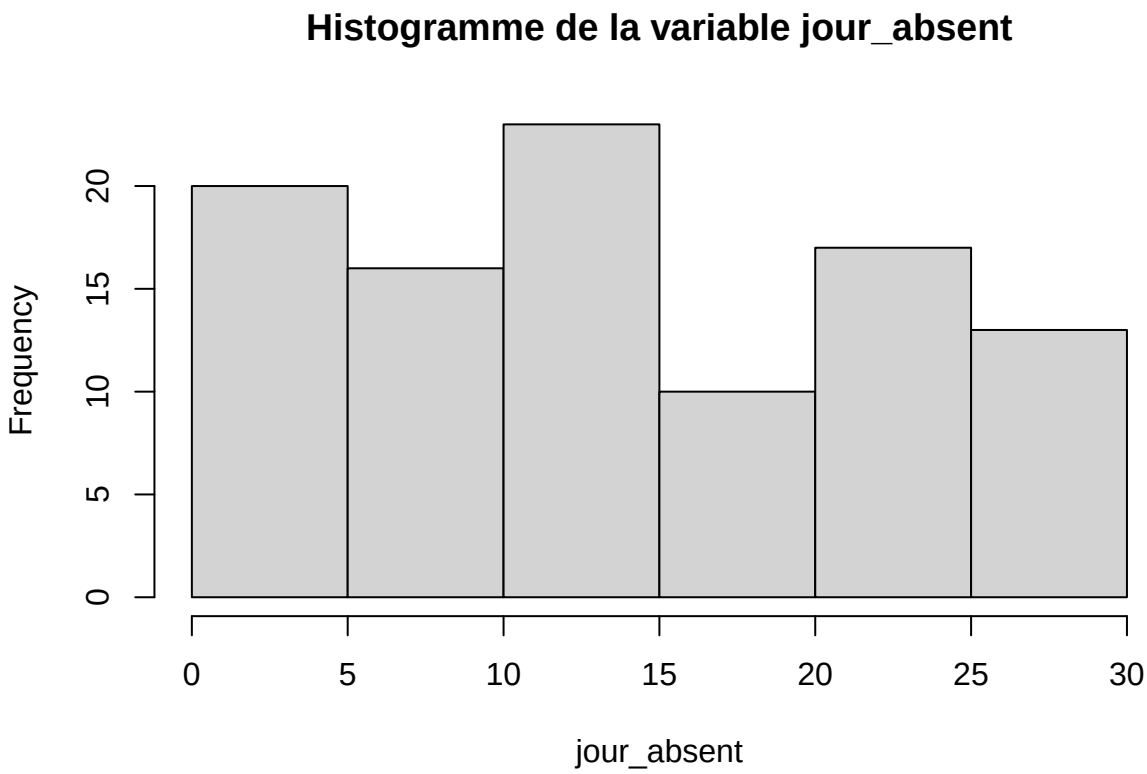


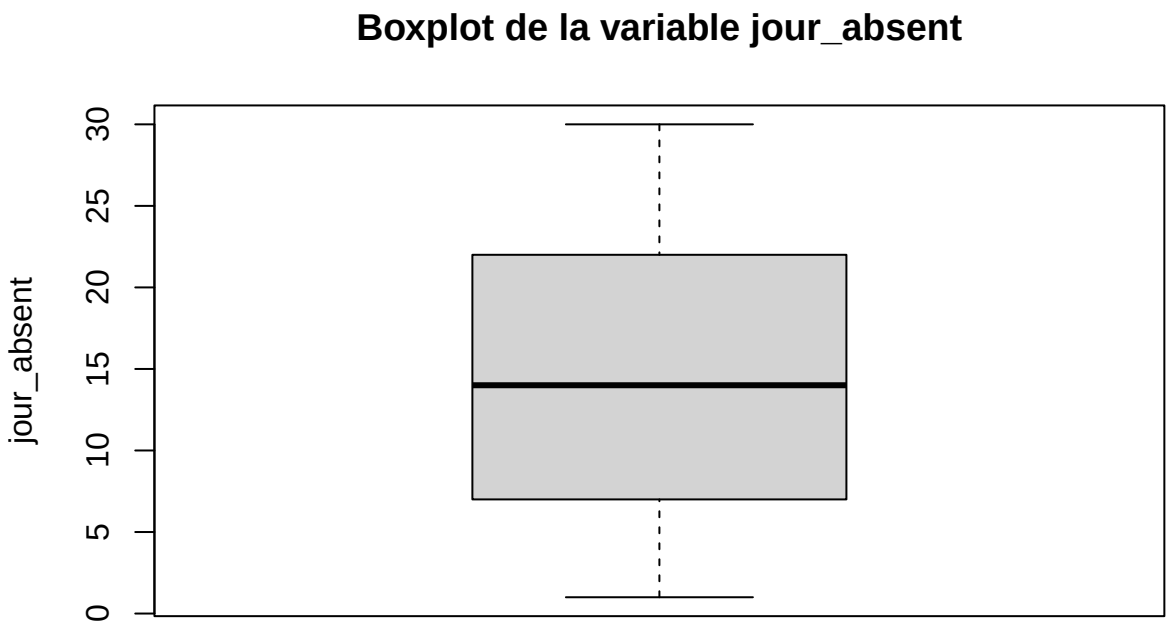

```
## Intervalle de confiance à 95%: [1011.53, 1216.51]
## function (x, ...)
## UseMethod("hist")
## <bytecode: 0x0000019a319062b8>
## <environment: namespace:graphics>
## function (x, ...)
## UseMethod("boxplot")
## <bytecode: 0x0000019a3194f068>
## <environment: namespace:graphics>
```

jour_absent

```
d.var.quant(df_Diop,"jour_absent")
```

```
## Tendances centrales:
## Moyenne: 14.26
## Médiane: 14
## Écart-type: 8.65
```


Histogramme de la variable jour_absent
Frequency
0
5
10
15
20
0
5
10
15
20
25
30
jour_absent




```
## Intervalle de confiance à 95%: [12.54, 15.99]
## function (x, ...)
## UseMethod("hist")
## <bytecode: 0x0000019a319062b8>
## <environment: namespace:graphics>
## function (x, ...)
## UseMethod("boxplot")
## <bytecode: 0x0000019a3194f068>
## <environment: namespace:graphics>
```

d.var.quali

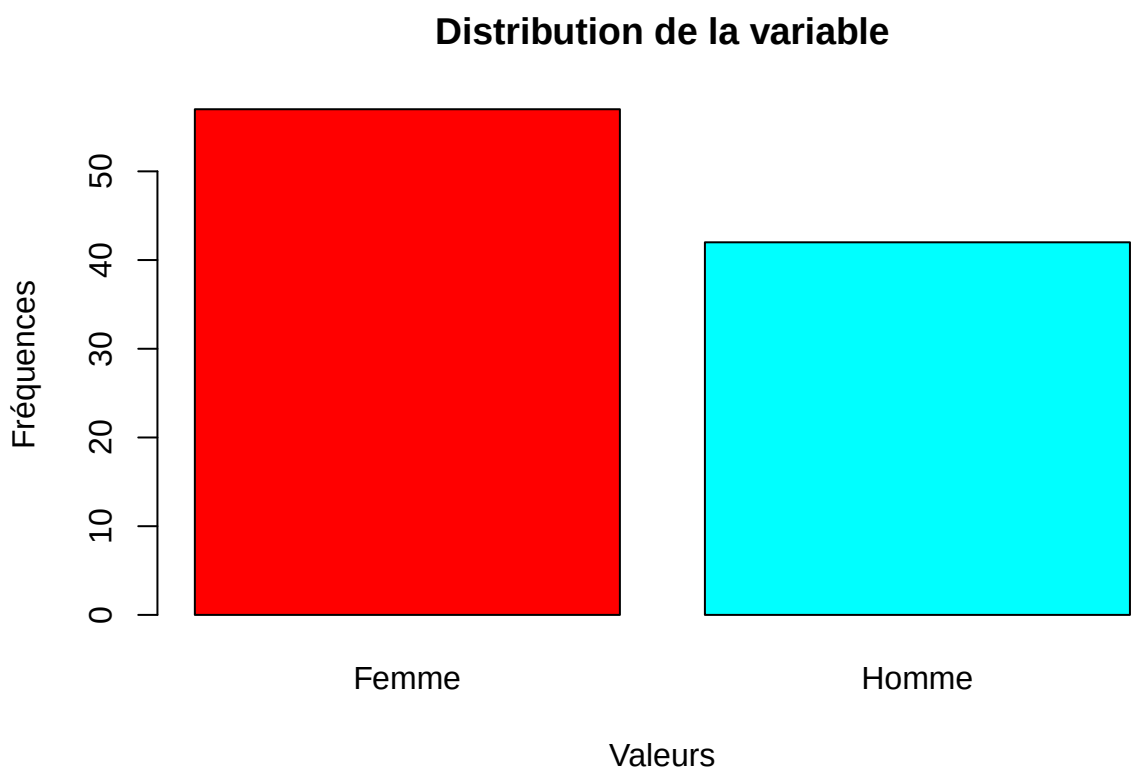
```
knitr::opts_chunk$set(echo=TRUE, error=FALSE, warning=FALSE, message=TRUE)
d.var.quali <- function(var){
  # Création d'un tableau de fréquences
  freq_table <- table(var)
  # Calcul des proportions
  prop_table <- prop.table(freq_table)
  # Affichage du tableau de fréquences et des proportions
  cat("Tableau de fréquences :\n")
  print(freq_table)
  cat("\nTableau des proportions :\n")
  print(prop_table)
  # Création d'un graphique en barres
```

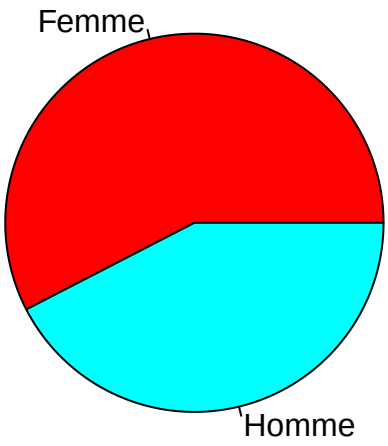
```
  barplot(freq_table, main="Distribution de la variable", xlab="Valeurs", ylab="Fréquences", col=rainbo
  # Création d'un diagramme en secteurs
  pie(freq_table, main="Répartition de la variable", col=rainbow(length(freq_table)))

}
```

Genre

```
genre<-df_Diop$genre
d.var.quali(genre)
```

```
## Tableau de fréquences :
## var
## Femme Homme
##    57    42
##
## Tableau des proportions :
## var
##     Femme     Homme
## 0.5757576 0.4242424
```



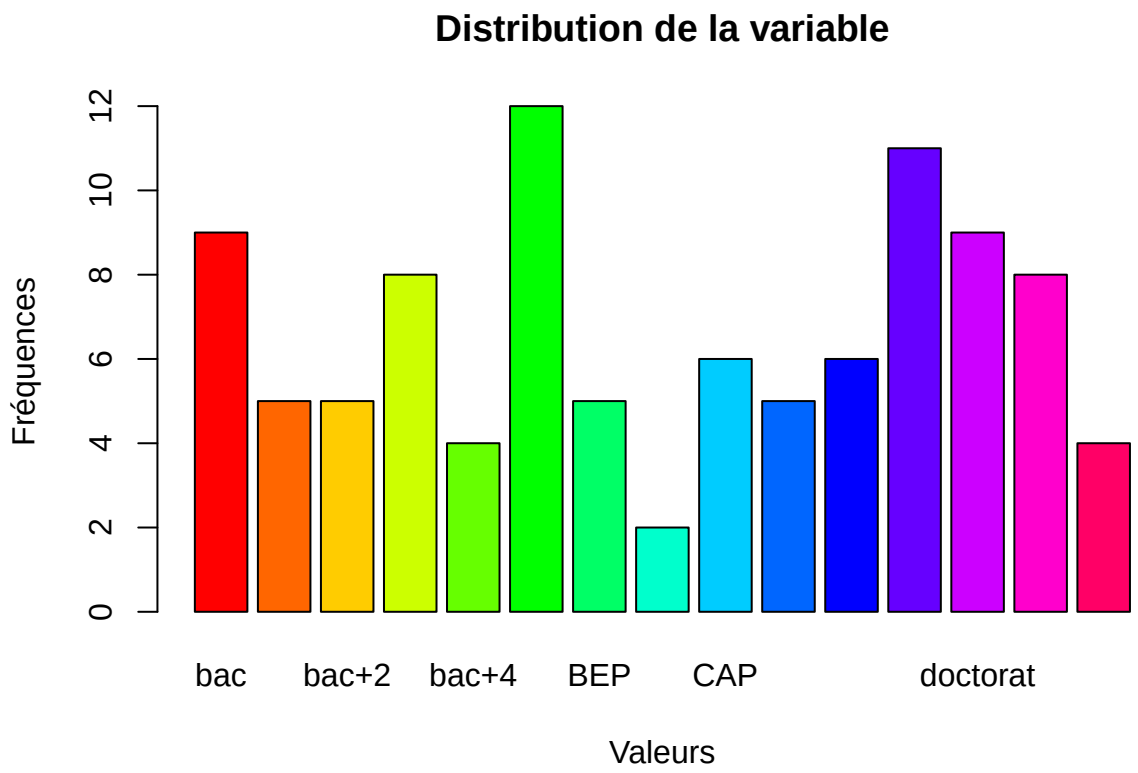


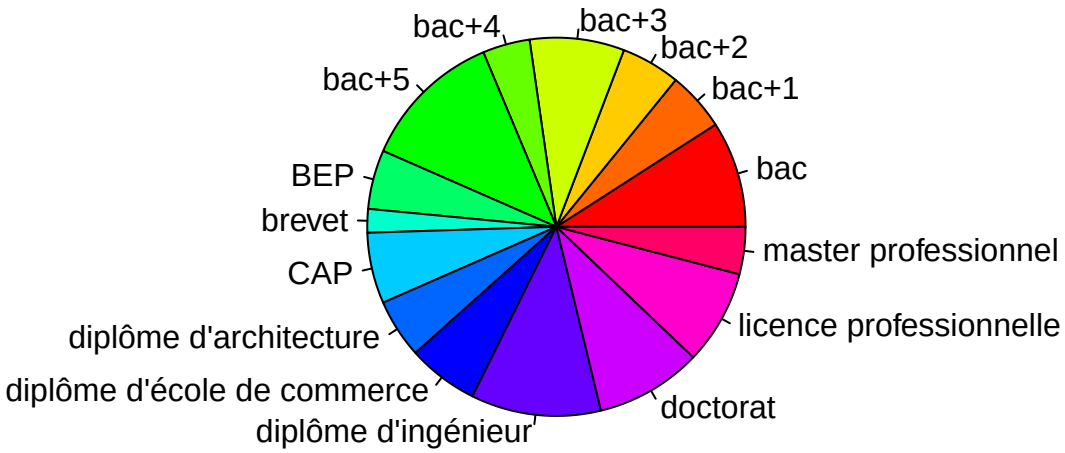
niveau_etude

```
niveau_etude<-df_Diop$niveau_etude
d.var.quali(niveau_etude)
```

```
## Tableau de fréquences :
## var
##                           bac                          bac+1
##                             9                              5
##                         bac+2                          bac+3
##                             5                              8
##                         bac+4                          bac+5
##                             4                             12
##                           BEP                         brevet
##                             5                              2
##                           CAP       diplôme d'architecture
##                             6                              5
## diplôme d'école de commerce          diplôme d'ingénieur
##                             6                             11
##                      doctorat      licence professionnelle
##                             9                              8
##         master professionnel
##                             4
##
```

```
## Tableau des proportions :
## var
##                          bac                         bac+1
##                   0.09090909                   0.05050505
##                        bac+2                        bac+3
##                   0.05050505                   0.08080808
##                        bac+4                        bac+5
##                   0.04040404                   0.12121212
##                          BEP                       brevet
##                   0.05050505                   0.02020202
##                          CAP       diplôme d'architecture
##                   0.06060606                   0.05050505
## diplôme d'école de commerce          diplôme d'ingénieur
##                   0.06060606                   0.11111111
##                     doctorat      licence professionnelle
##                   0.09090909                   0.08080808
##          master professionnel
##                   0.04040404
```

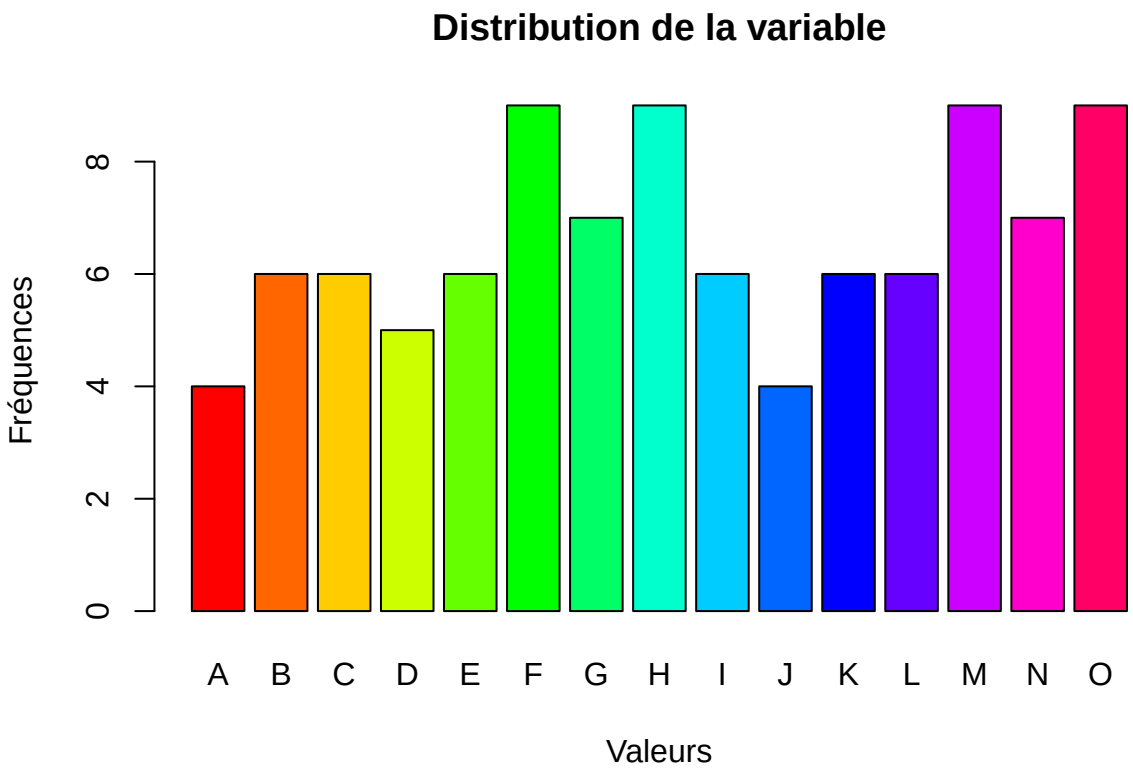


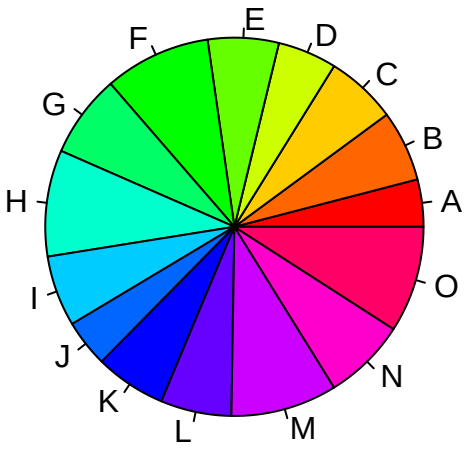


lettre_preferee

```
lettre_preferee<-df_Diop$lettre_preferee
d.var.quali(lettre_preferee)
```

```
## Tableau de fréquences :
## var
## A B C D E F G H I J K L M N O
## 4 6 6 5 6 9 7 9 6 4 6 6 9 7 9
##
## Tableau des proportions :
## var
##         A         B         C         D         E         F         G
## 0.04040404 0.06060606 0.06060606 0.05050505 0.06060606 0.09090909 0.07070707
##         H         I         J         K         L         M         N
## 0.09090909 0.06060606 0.04040404 0.06060606 0.06060606 0.09090909 0.07070707
##         O
## 0.09090909
```

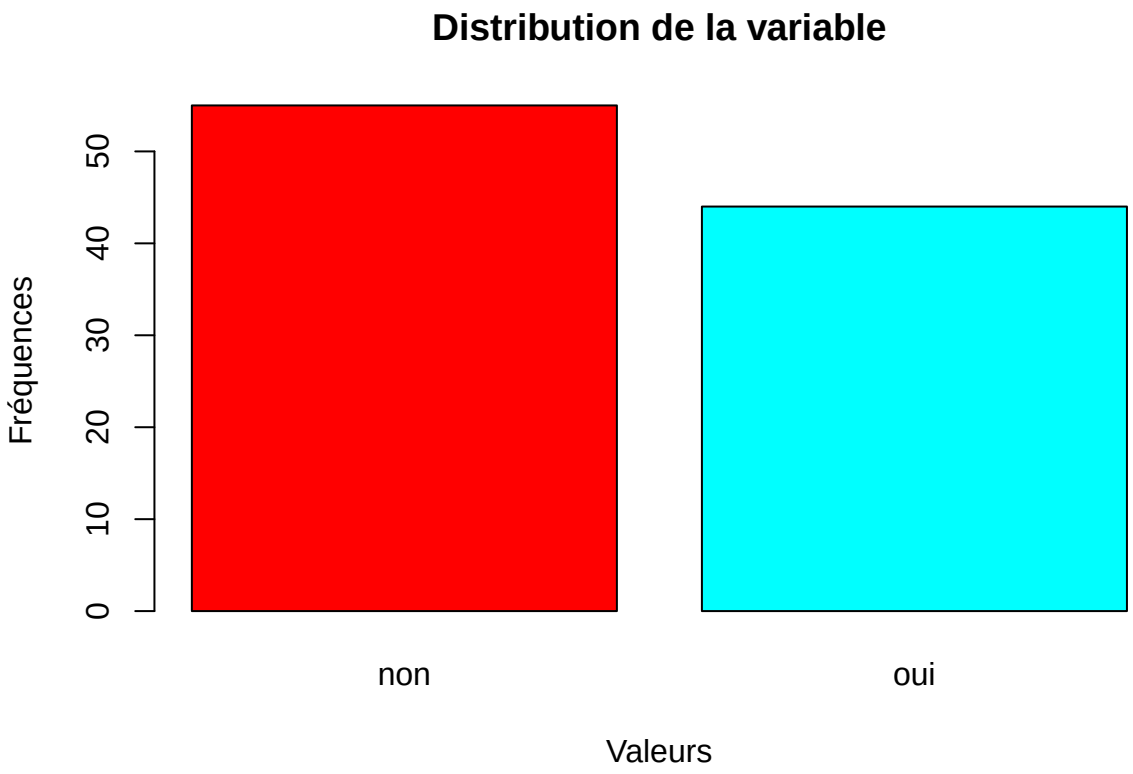

Distribution de la variable
Fréquences
0
2
4
6
8
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
Valeurs

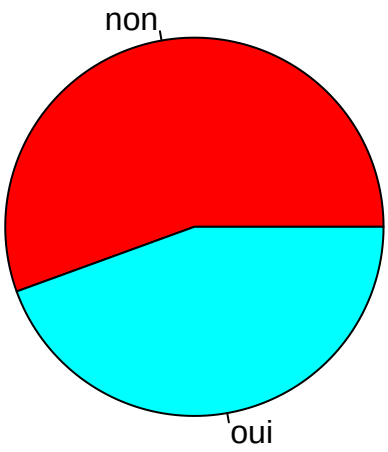


voyage_etude

```
voyage_etude<-df_Diop$voyage_etude
d.var.quali(voyage_etude)
```

```
## Tableau de fréquences :
## var
## non oui
##  55  44
##
## Tableau des proportions :
## var
##       non       oui
## 0.5555556 0.4444444
```


Distribution de la variable
Fréquences
0
10
20
30
40
50
non
oui
Valeurs



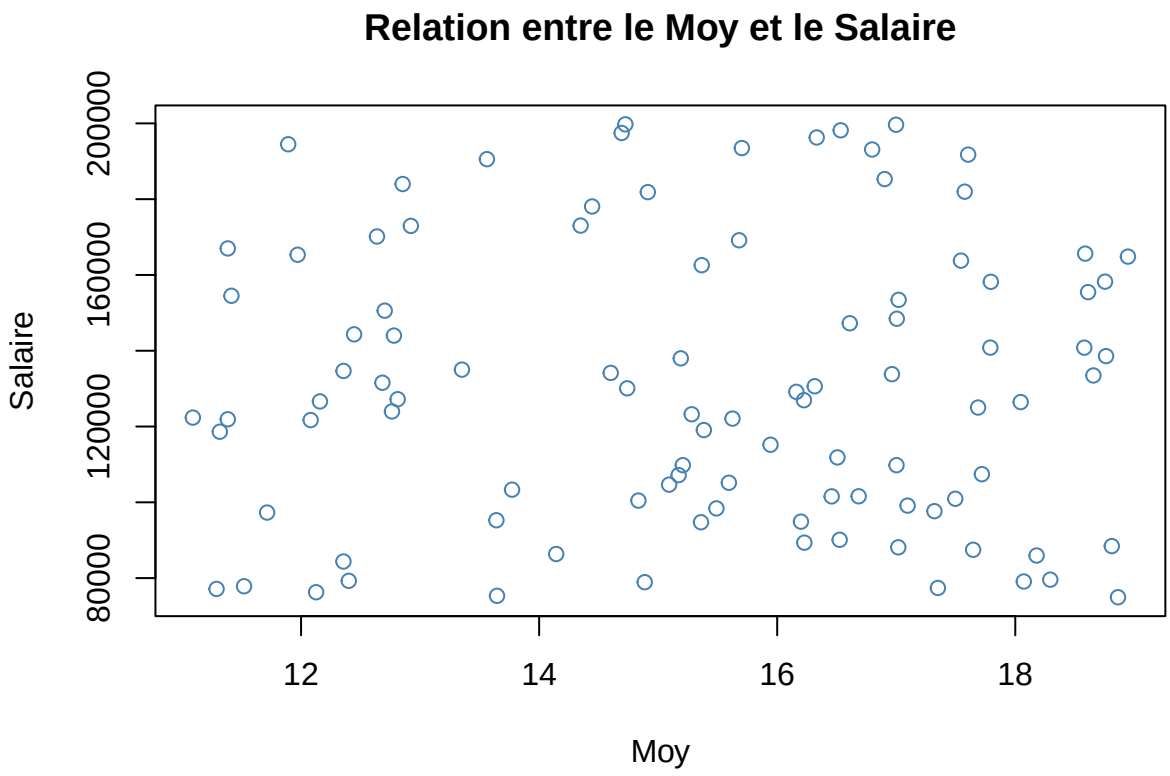
Question 6

variables quantitatives

Comme les variables “moy” et “salaire” dans la base de données “df_Diop” sont toutes les deux quantitatives, nous pouvons examiner leur relation à l’aide d’un graphique de dispersion et d’un coefficient de corrélation.

Tout d’abord, nous allons tracer le graphique de dispersion à l’aide de la fonction “plot” de R :

```
plot(df_Diop$moy, df_Diop$salaire, main="Relation entre le Moy et le Salaire", xlab="Moy", ylab="Salair
```



Ensuite, nous allons calculer le coefficient de corrélation entre les deux variables à l'aide de la fonction "cor" de R :

```
cor(df_Diop$moy, df_Diop$salaire)
```

```
## [1] 0.001833772
```

Le coefficient de corrélation est de 0.001833772, ce qui suggère qu'il n'y a pas de relation entre les deux variables.

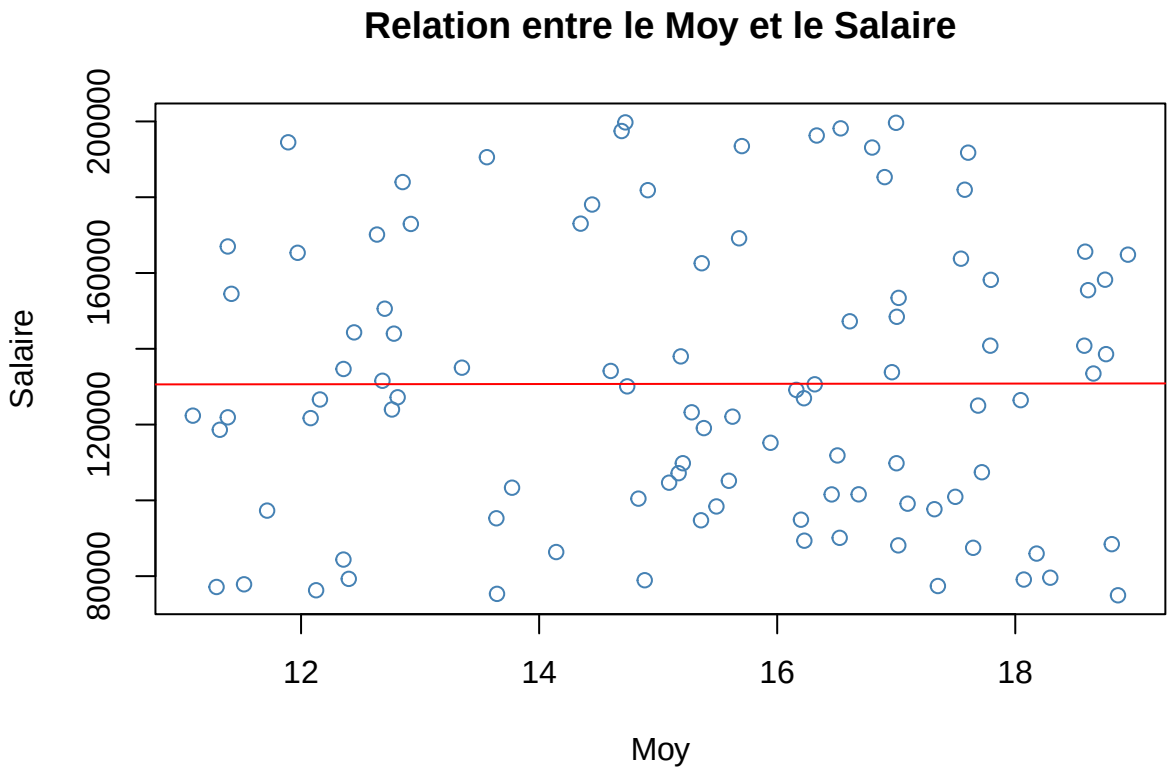
passons la la rg

```
model <- lm(salaire ~ moy, data=df_Diop)
summary(model)
```

```
##
## Call:
## lm(formula = salaire ~ moy, data = df_Diop)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55892 -30965  -3815  29613  69019
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 130287.3    25316.8   5.146 1.39e-06 ***
## moy             29.5     1633.3   0.018    0.986
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37180 on 97 degrees of freedom
## Multiple R-squared:  3.363e-06,  Adjusted R-squared:  -0.01031
## F-statistic: 0.0003262 on 1 and 97 DF,  p-value: 0.9856
```

```
plot(df_Diop$moy, df_Diop$salaire, main="Relation entre le Moy et le Salaire", xlab="Moy", ylab="Salaire
abline(model, col="red")
```



Variables qualitatives

nous souhaitons décrire la liaison entre le genre et le niveau d'étude dans la base de données df_Diop. Nous pouvons donc créer un tableau de contingence qui présente le nombre d'individus en fonction de leur genre et de leur niveau d'étude.

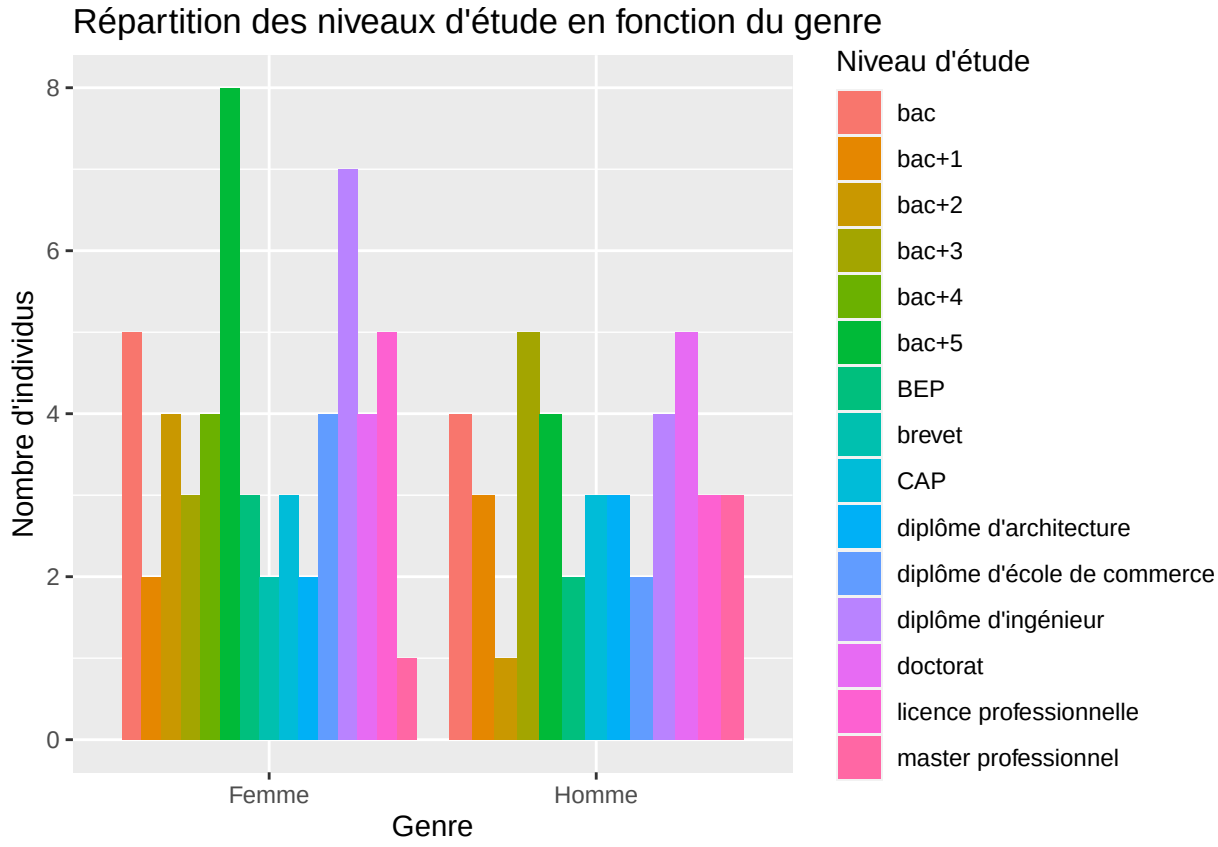
```
table(df_Diop$genre, df_Diop$niveau_etude)
```

```
##
##         bac bac+1 bac+2 bac+3 bac+4 bac+5 BEP brevet CAP diplôme d'architecture
##   Femme   5     2     4     3     4     8   3      2   3                      2
##   Homme   4     3     1     5     0     4   2      0   3                      3
```

```
##
##          diplôme d’école de commerce diplôme d’ingénieur doctorat
##   Femme                            4                   7        4
##   Homme                            2                   4        5
##
##          licence professionnelle master professionnel
##   Femme                        5                    1
##   Homme                        3                    3
```

avec grapgique:

```
ggplot(df_Diop, aes(x=genre, fill=niveau_etude)) +
  geom_bar(position="dodge") +
  ggtitle("Répartition des niveaux d'étude en fonction du genre") +
  xlab("Genre") +
  ylab("Nombre d'individus") +
  scale_fill_discrete(name="Niveau d'étude")
```



liason entre genre et salaire

Tout d’abord, nous pouvons créer une table de contingence pour visualiser la distribution des salaires en fonction du genre .

```
table(df_Diop$genre, df_Diop$salaire)
```

```
##
##         74952 75332 76279 77142 77402 77851 78926 79107 79282 79591 84382 85961
##   Femme     1     0     1     0     0     0     0     1     0     0     0     1
##   Homme     0     1     0     1     1     1     1     0     1     1     1     0
##
##         86367 87481 88122 88445 89361 90116 94743 94901 95275 97298 97668 98400
##   Femme     1     1     1     0     1     1     0     0     0     1     1     1
##   Homme     0     0     0     1     0     0     1     1     1     0     0     0
##
##         99130 100457 100935 101574 101589 103341 104668 105167 107176 107423
##   Femme     0      1      1      0      0      1      1      0      1      1
##   Homme     1      0      0      1      1      0      0      1      0      0
##
##         109799 109803 111850 115203 118627 119071 121705 121913 122093 122356
##   Femme      1      1      0      1      1      1      0      0      1      0
##   Homme      0      0      1      0      0      0      1      1      0      1
##
##         123240 123982 125012 126447 126621 126951 127203 129149 130083 130627
##   Femme      1      1      1      1      0      1      1      0      0      1
##   Homme      0      0      0      0      1      0      0      1      1      0
##
##         131563 133474 133818 134154 134672 135014 137991 138585 140816 140824
##   Femme      1      1      0      1      1      1      1      0      0      1
##   Homme      0      0      1      0      0      0      0      1      1      0
##
##         143997 144325 147254 148445 150577 153434 154500 155470 158187 158231
##   Femme      1      1      1      0      0      0      0      0      1      1
##   Homme      0      0      0      1      1      1      1      1      0      0
##
##         162577 163775 164843 165331 165625 167000 169153 170149 172948 173027
##   Femme      0      0      0      0      0      1      1      1      1      1
##   Homme      1      1      1      1      1      0      0      0      0      0
##
##         178070 181854 181972 183992 185300 190557 191753 193110 193487 194486
##   Femme      1      1      0      0      0      1      1      1      1      0
##   Homme      0      0      1      1      1      0      0      0      0      1
##
##         196278 197466 198170 199636 199741
##   Femme      1      1      0      0      1
##   Homme      0      0      1      1      0
```

utilisons un test d'indépendance du Chi-squared pour déterminer s'il existe une association significative entre le genre et le salaire :

```
chisq.test(df_Diop$genre, df_Diop$salaire)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_Diop$genre and df_Diop$salaire
## X-squared = 99, df = 98, p-value = 0.4527
```

Pour explorer la relation linéaire entre le genre et le salaire, nous pouvons tracer un diagramme de dispersion:

```
library(ggplot2)
ggplot(df_Diop, aes(x = genre, y = salaire)) +
  geom_point(color = "steelblue") +
  ggtitle("Relation entre le genre et le salaire") +
  xlab("Genre") +
  ylab("Salaire")
```