1-

→read CHD :

```
> library(readxl)
> data <- read_excel("CHD.xlsx")
> head(data)
# A tibble: 6 x 7
    sbp   ldl adiposity famhist obesity   age   chd
  <dbl> <dbl>     <dbl> <chr>     <dbl> <dbl> <dbl>
1   160   573      2311 Present     253    52     1
2   144   441      2861 Absent     2887    63     1
3   118   348      3228 Present    2914    46     0
4   170   641      3803 Present    3199    58     1
5   134    35      2778 Present    2599    49     1
6   132   647      3621 Present    3077    45     0
>
```

→dimensions :

```
> dim(data)
[1] 462   7
>
```

2-

→tester si ils sont qualitatives :

```
> is.factor(data$chd)
[1] FALSE
> is.factor(data$famhist)
[1] FALSE
>
```

On remarque que ils ne sont pas de type qualitatives.

→on les transforme à ce type :

```
> data$chd = as.factor(data$chd)
> data$famhist = as.factor(data$famhist)
> is.factor(data$chd)
[1] TRUE
> is.factor(data$famhist)
[1] TRUE
>
```

3-

→on tape la cmd suivante :

Après cette map on remarque que la base de données ne contient pas des données manquantes.

→Mais si on a obtenu des donnée manquants et on veut les supprimer on tape :

```
data = na.omit(data)
```

4-

→var age

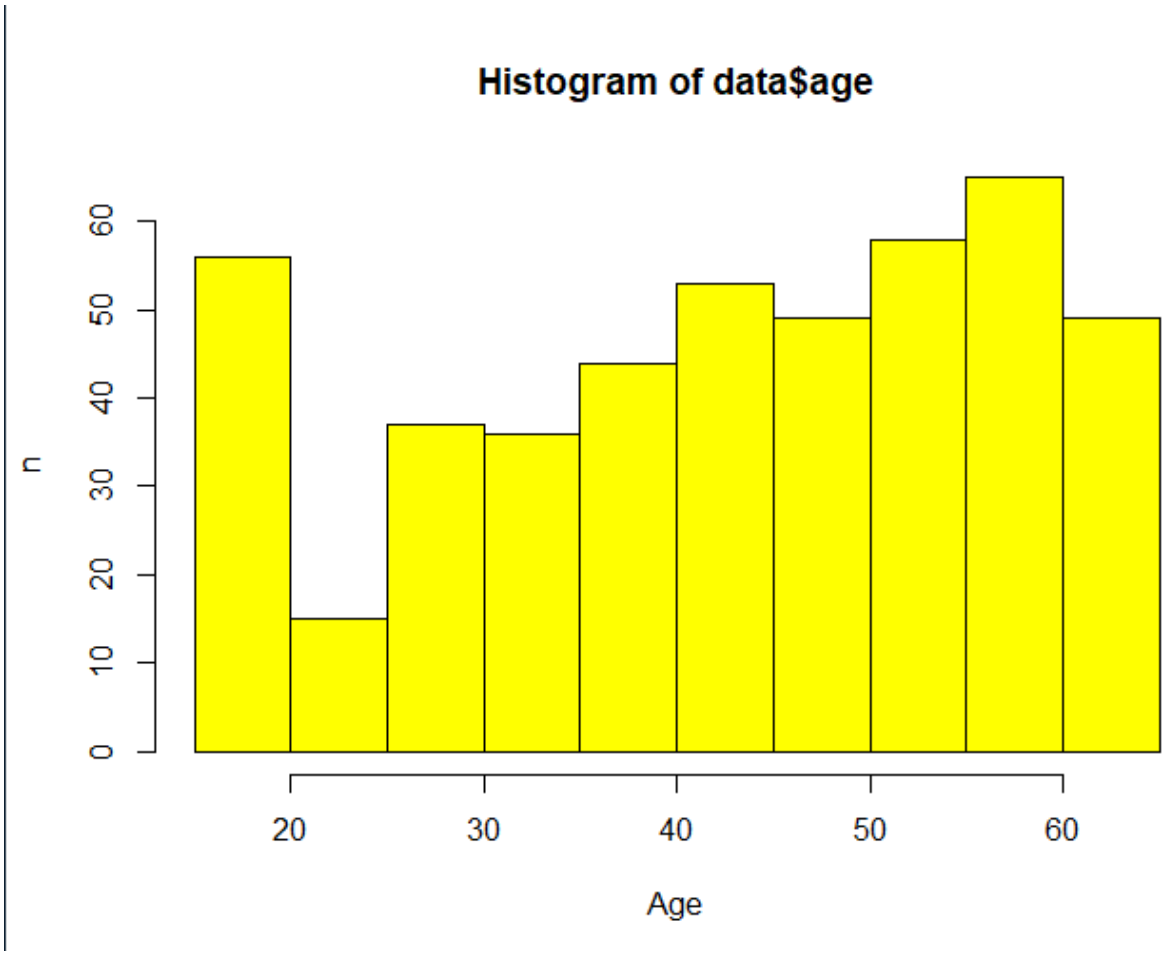
→indicateurs statistiques

```
> mean(data$age,na.rm = TRUE)
[1] 42.81602
> median(data$age,na.rm = TRUE)
[1] 45
> sd(data$age,na.rm = TRUE)
[1] 14.60896
> var(data$age,na.rm = TRUE)
[1] 213.4216
> quantile(data$age,na.rm = TRUE)
  0%  25%  50%  75% 100%
  15   31   45   55   64
> IQR(data$age,na.rm = TRUE)
[1] 24
> range(data$age,na.rm = TRUE)
[1] 15 64
> library(e1071)
> skewness(data$age,na.rm = TRUE)
[1] -0.379259
> kurtosis(data$age,na.rm = TRUE)
[1] -1.026793
>
```

→représentation graphique :

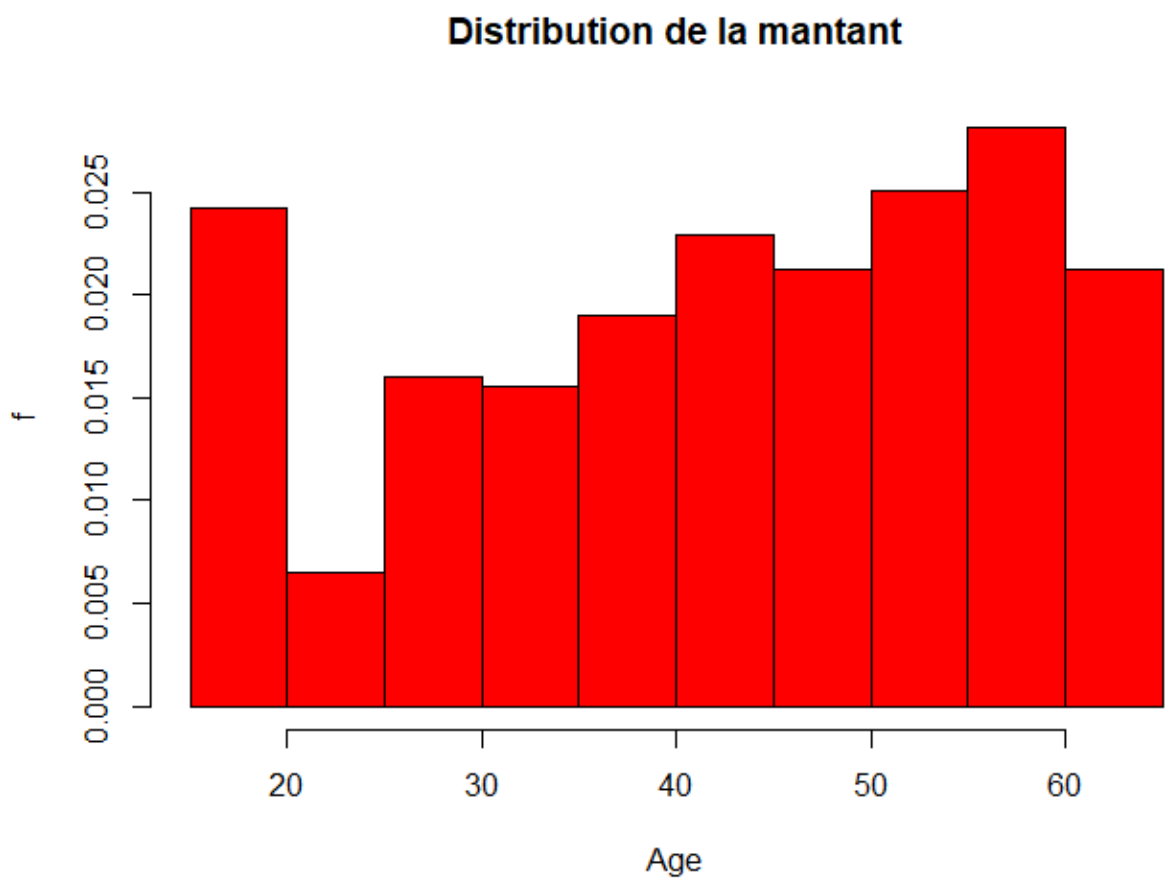
-hist des effectifs

```
> hist(data$age,col="yellow",xlab = "Age",ylab = "n")
>
```



- hist des frequences

```
> hist(data$age,col="red",xlab = "Age",ylab = "f",main = "Distribution de la mantant",p
robability = T)
>
```



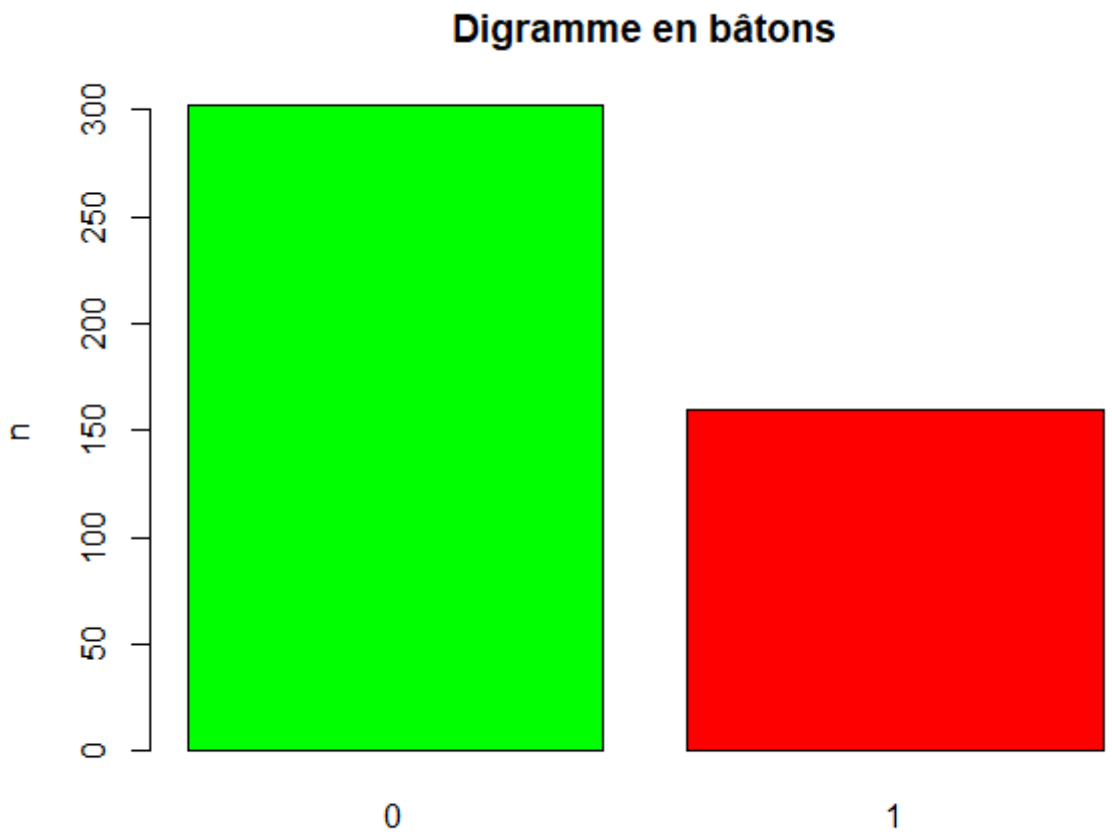
→var chd

-indicateurs statistiques

```
> table(data$chd)

  0   1
302 160
> library(questionr)
> freq(data$chd)
    n    % val%
0 302 65.4 65.4
1 160 34.6 34.6
>
```

-représentation graphique

```
> barplot(table(data$chd),col=c("green","red"),main="Digramme en bâtons",ylab ="n")
>
```

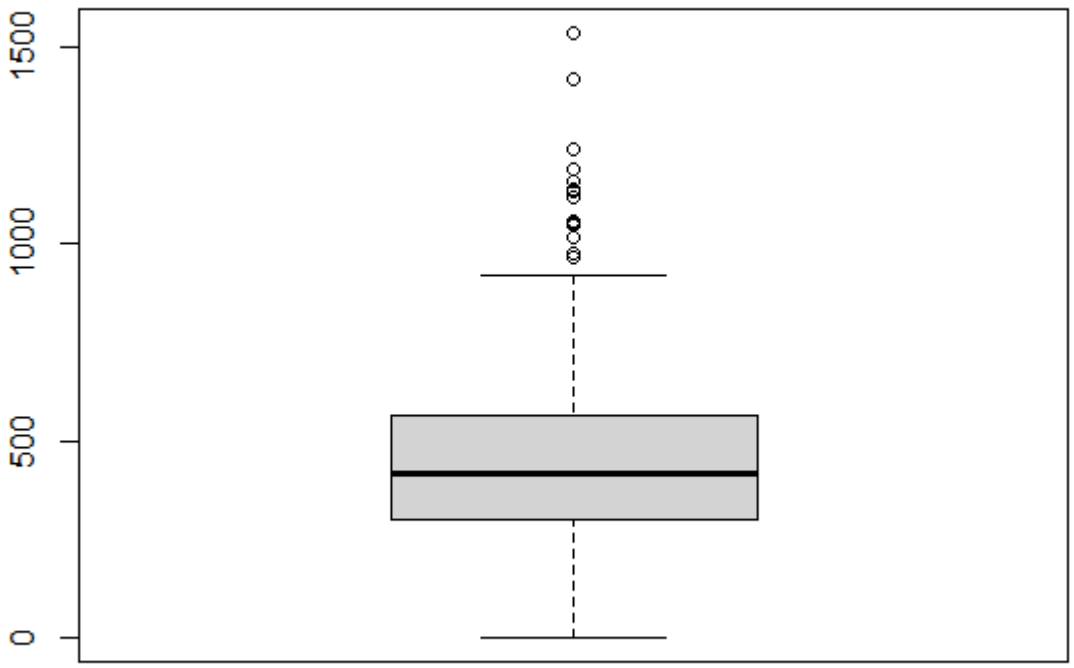


-diagramme en secteurs :

```
> pie(table(data$chd),col=c("blue","black"),main="Digramme en secteur")
>
```

5-on trace la boite à moustache de cette var par la cmd :

```
> boxplot(data$ldl)
```



Après cette figure on voie bien qu'elle contient des données aberrantes.

-on les affiche par la cmd :

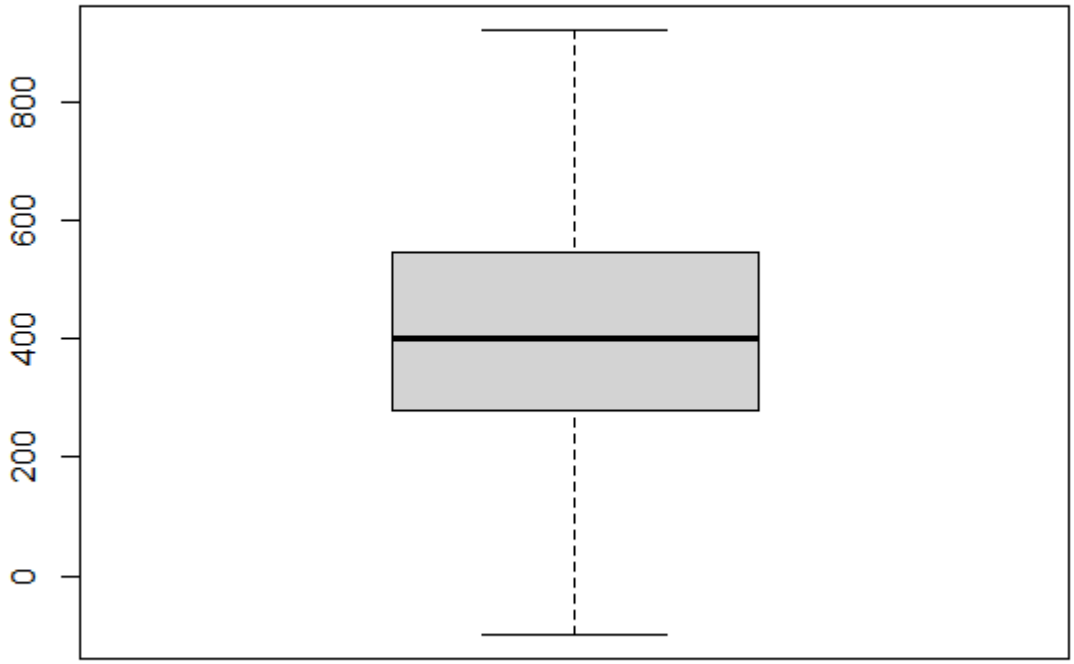
```
> boxplot.stats(data$ldl)$out
 [1] 1533 1242  965 1132 1058 1117 1053 1189 1141 1049 1019  978 1416 1161
>
```

Et on les remplace par a et b par la cmd :

```
> q=quantile(data$ldl,probs=c(0.25,0.75),na.rm=TRUE)
> a=q[2]+1.5*IQR(data$ldl,na.rm=TRUE)
> b=q[1]-1.5*IQR(data$ldl,na.rm=TRUE)
> data$ldl[data$ldl>a]=b
> data$ldl[data$ldl<b]=a
>
```

Donc :

```
> boxplot.stats(data$ldl)$out
numeric(0)
> boxplot(data$ldl)
>
```



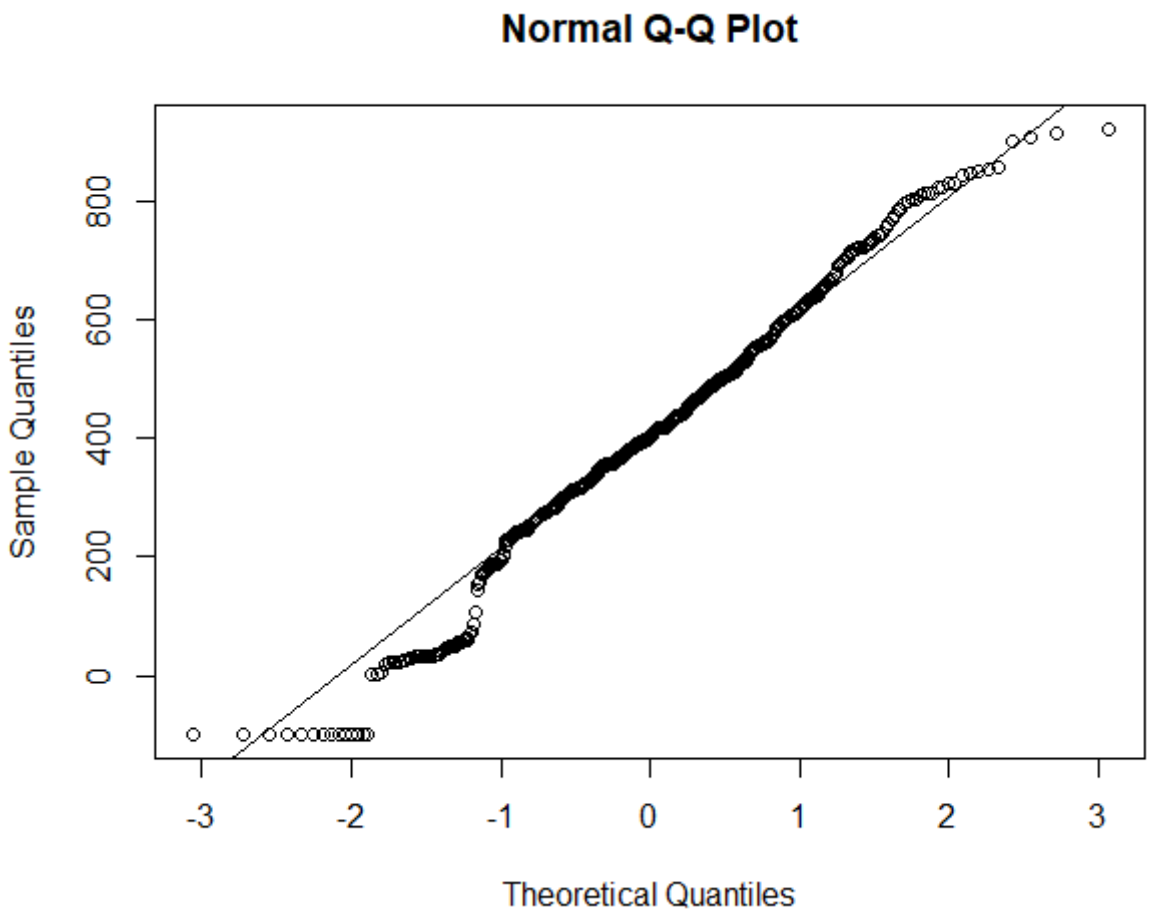
Alors il n’ y a pas des données manquants

6-

➔ var ldl

→Méthode graphique :

```
> qqnorm(data$ldl)
> qqline(data$ldl)
```



après cette figure on remarque il n'y a pas de normalité de ce var car les données ne sont pas linéaire.

→méthode basée sur les tests :

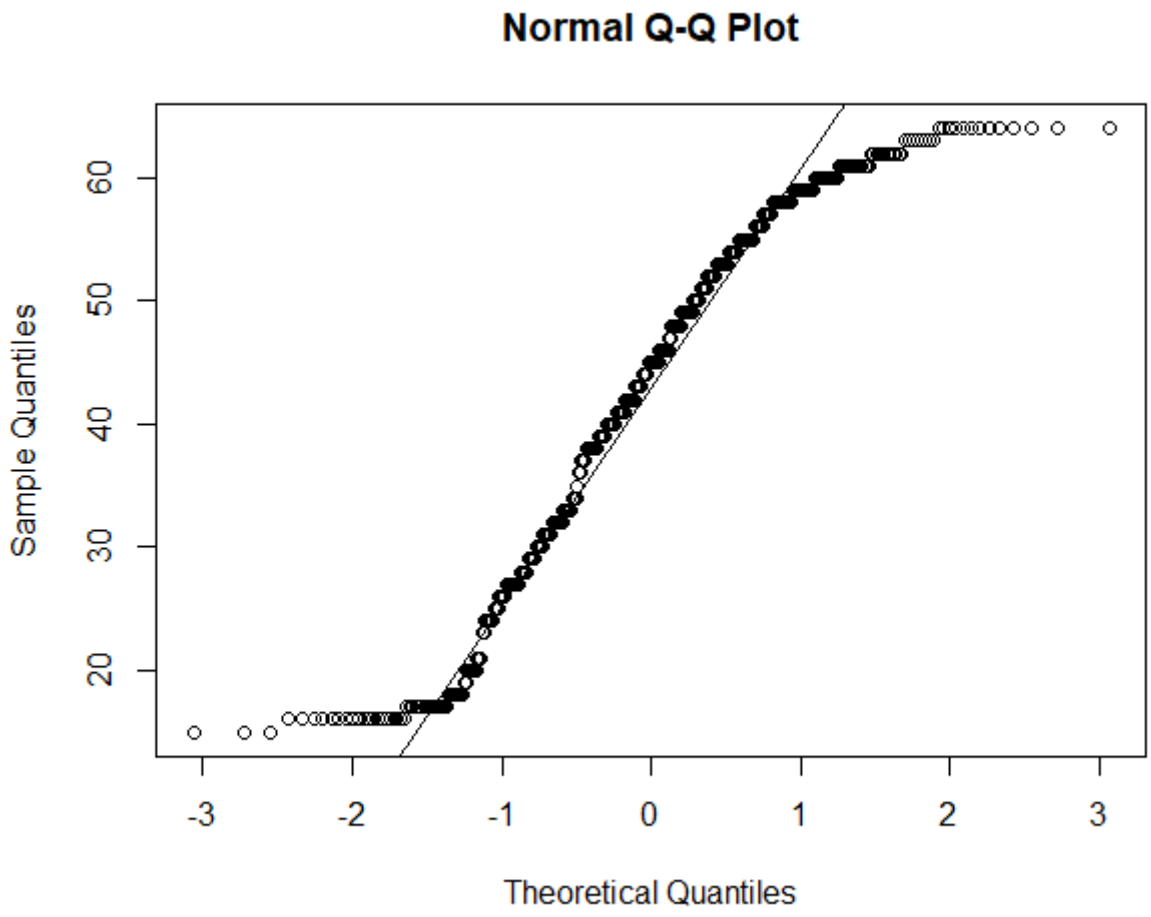
```
> shapiro.test(data$ldl)

        Shapiro-Wilk normality test

data:  data$ldl
W = 0.98553, p-value = 0.0001503
```

On a p-value < 0.05 alors on rejette H0 (pas de normalité de var ldl)

➔ var age

→Méthode graphique :

```
> qqnorm(data$age)
> qqline(data$age)
>
```



après cette figure on remarque il n'y a pas de normalité de ce var car les données ne sont pas linéaire.

→méthode basée sur les tests :

```
> shapiro.test(data$age)

        Shapiro-Wilk normality test

data:  data$age
W = 0.93705, p-value = 4.595e-13
```

Après ce test de shapirio on remarque que P-value <<<< 0.05

Donc on rejette la normalité.

7-

On utilise ici le test de corrélation entre deux var non normaux distribuées

```
> cor.test(data$ldl,data$age,method="spearman")

        Spearman's rank correlation rho

data:  data$ldl and data$age
S = 12349768, p-value = 1.172e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2436735

Warning message:
In cor.test.default(data$ldl, data$age, method = "spearman") :
  Cannot compute exact p-value with ties
>
```

P-value <<< 0.05 donc pas de corrélation entre ces deux variables.

8-

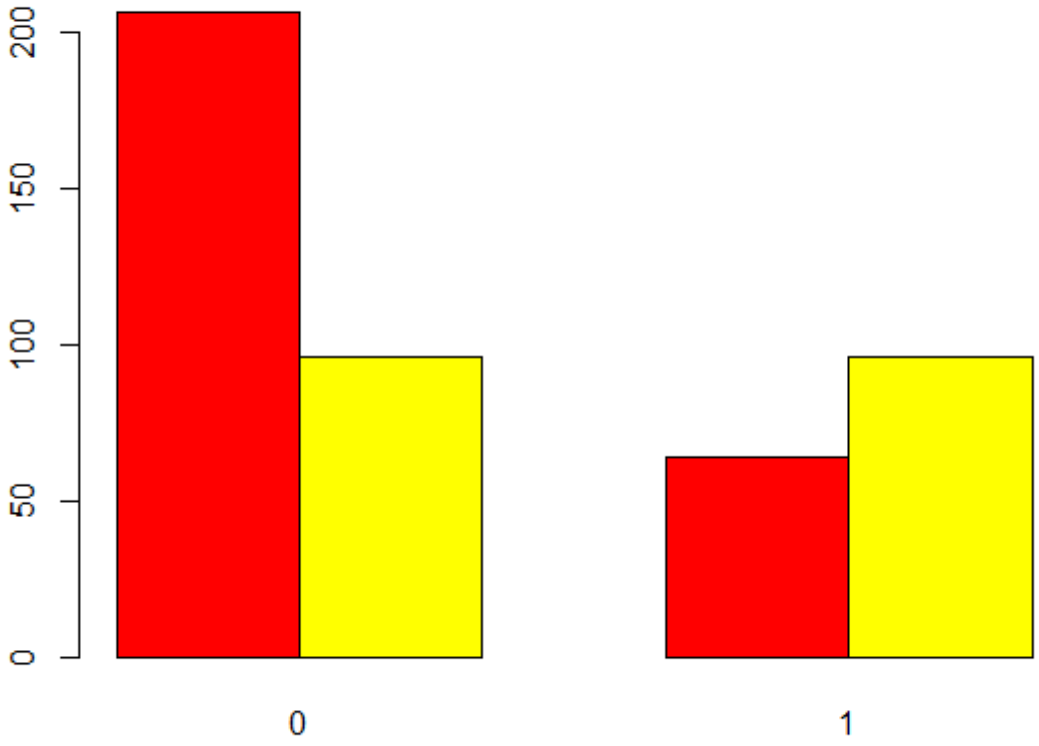
→Indicateurs statistique :

tableau croisé de ces deux var quali :

```
> table(data$famhist, data$chd)

            0   1
  Absent  206  64
  Present  96  96
```

un diagramme en bâton juxtaposé pour deux var quali à 2 modalités :

```
> barplot(table(data$famhist, data$chd) ,beside = T , col=c("red","yellow"))
>
```



9-

On fait le test de kh-deux d'indépendance entre deux var quali :

```
> chisq.test(data$famhist, data$chd)

        Pearson's Chi-squared test with Yates' continuity correction

data:  data$famhist and data$chd
X-squared = 33.123, df = 1, p-value = 8.653e-09

>
```

On voit que p-value < 0.05 danc il n'y pas d'indépendance entre ces deux variables

Alors on peut dire qu'il y a une relation significative entre ces deux variables

10-

On fait le test d'ajustement de khideux:

11-

→indicateurs statistiques

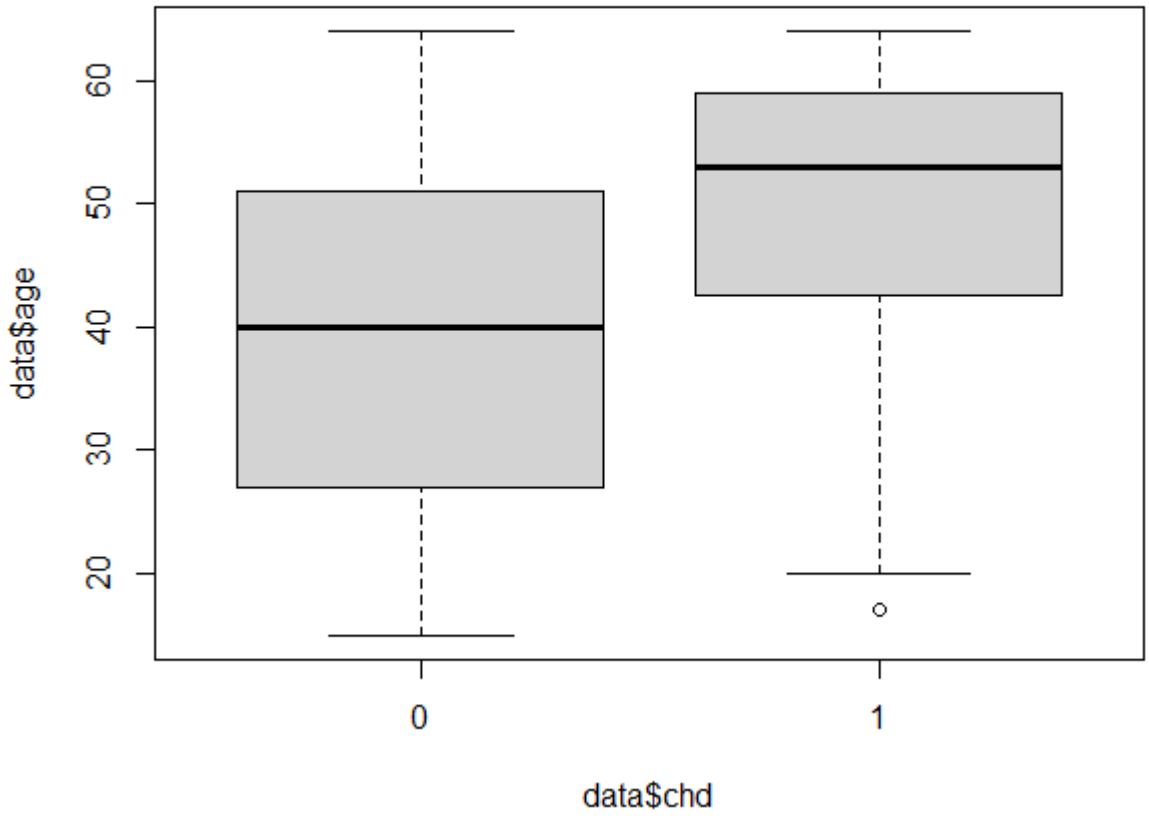
Croisement de la variable âge avec la variable chd :

```
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

> data%>%group_by(data$chd)%>%summarise(mean(data$age),median(data$age),sd(data$age))
# A tibble: 2 x 4
  `data$chd` `mean(data$age)` `median(data$age)` `sd(data$age)`
  <fct>                <dbl>              <dbl>          <dbl>
1 0                     42.8                 45           14.6
2 1                     42.8                 45           14.6
>
```

On remarque ici que les médians sont égaux et aussi les écarts-types et les moyenne. On conclue que l'âge n'influence pas sur la maladie.

→représentation graphique :

```
> boxplot(data$age~data$chd)
>
```



Ici on remarque que les gens les plus âgées ont plus de chance d'avoir le maladie en comparaissant avec les personnes les moins âgées.

```
> t.test(data$age~data$chd, mu =0, alternative = "two.sided", paired = FALSE, conf.leve
l = 0.95, var.equal = TRUE)

        Two Sample t-test

data:  data$age by data$chd
t = -8.6215, df = 460, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equ
al to 0
95 percent confidence interval:
 -14.046887  -8.832004
sample estimates:
mean in group 0 mean in group 1
       38.85430        50.29375

>
```

p-value < 0.05 : donc il n'y a pas d'association significative.

14-

1. Variable "ldl" : Après l'examen d'une figure, il a été observé que la variable "ldl" ne présente pas de distribution normale car les données ne suivent pas une tendance linéaire.
2. Variable "age" : Une autre figure a été examinée et il a été constaté que la variable "age" ne présente pas de distribution normale car les données ne suivent pas une tendance linéaire.
3. Corrélation entre "age" et "ldl" : Il a été remarqué qu'il n'y a pas de corrélation entre les variables "age" et "ldl". Cela indique qu'il n'y a pas de relation linéaire évidente entre l'âge et le taux de cholestérol LDL.
4. Indépendance entre "chd" et "famhist" : Il a été observé qu'il n'y a pas d'indépendance entre les variables "chd" (présence ou absence d'une maladie cardiaque) et "famhist" (antécédents familiaux de maladie cardiaque). Cela suggère qu'il existe une association entre ces deux variables, ce qui signifie que les antécédents familiaux peuvent influencer la prédisposition à la maladie cardiaque.

Ces observations soulignent des caractéristiques importantes de la base de données, notamment l'absence de normalité dans les variables "ldl" et "age", l'absence de corrélation entre "age" et "ldl", ainsi que la présence d'une association entre "chd" et "famhist". Ces informations peuvent orienter les analyses et les interprétations ultérieures des données.