# PRA2

Toni Vanrell i Guillem Mir

31/5/2021

## 1. Descripció

El subdataset escollit és un conegut dataset de Kaggle amb dades relacionades amb malalties cardíaques. Les variables que inclou el dataset són:

1. Edat: Edat de la persona.
2. Sexe: Sexe biològic de la persona.
3. Dolor al pit: Tipus de dolor al pit (Asimptòmatic, típic, atípic o sense dolor).
4. Pressió sanguínea: Pressió sanguínea en descans.
5. Colesterol: Concentració de colesterol en sang en mg/dl.
6. Sucre a la sang: Classificació segons diabètic o no diabètic.
7. Electrocardiograma: Anomalis en la lectura de l'electrocardiograma.
8. Freqüència cardíaca màxima.
9. Angina induïda: Presència d'angina de pit induïda per exercici.
10. Depressió del segment ST: Presència del segment ST en l'electrocardiograma.
11. Pendent del segment ST: Pendent del segment ST.
12. Número de vasos majors: Número de vasos colorejats durant la revisió per fluoroscòpia.
13. Talassèmia: Presència de la condició hereditària coneguda com Talassèmia.
14. Target: Presència o no de malaltia cardíaca.

Aquest dataset ens permet estudiar quins factors tenen influència directa sobre l'aparició de malalties coronàries. La intenció del nostre treball és visualitzar possibles relacions que puguin derivar en mesures preventives.

## 2.

Integració i selecció de les dades d'interès.

```
# Carreguem el fitxer "SFO.csv"
data <- read.csv('heart.csv')

colnames(data) <- c('age', 'sex', 'chestPain', 'bloodPressure',
                    'cholesterol', 'bloodSugar', 'restecg',
                    'maxHeartRate', 'indAngina', 'stDepression',
                    'stSlope', 'numVessels', 'scintigraphy', 'target')
# Recodifiquem i seleccionem variables d'interés per la nostra anàlisi
data2 <- data %>% mutate(
  sex = if_else(sex == 1, "Male", "Female"),
```

```
  bloodSugar = ifelse(bloodSugar == 1, '> 120 mg/dl', '< 120 mg/dl'),
  chestPain = ifelse(chestPain == 1,'Typical Angina',
                     ifelse(chestPain == 2, 'Atypical Angina',
                            ifelse(chestPain == 3,'Non-Anginal Pain', 'Asymptomatic'))),
  indAngina = ifelse(indAngina == 0,'No','Yes'),
  target = if_else(target == 1, "YES", "NO")
) %>% mutate_if(is.character, as.factor)

data2 <- data2[,c(1:6,8,9,14)]
summary(data2)
```

```
##       age            sex                 chestPain   bloodPressure
##  Min.   :29.00   Female: 96   Asymptomatic    :143   Min.   : 94.0
##  1st Qu.:47.50   Male  :207   Atypical Angina : 87   1st Qu.:120.0
##  Median :55.00                Non-Anginal Pain: 23   Median :130.0
##  Mean   :54.37                Typical Angina  : 50   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##   cholesterol          bloodSugar   maxHeartRate   indAngina target
##  Min.   :126.0   < 120 mg/dl:258   Min.   : 71.0   No :204   NO :138
##  1st Qu.:211.0   > 120 mg/dl: 45   1st Qu.:133.5   Yes: 99   YES:165
##  Median :240.0                     Median :153.0
##  Mean   :246.3                     Mean   :149.6
##  3rd Qu.:274.5                     3rd Qu.:166.0
##  Max.   :564.0                     Max.   :202.0
```

Les variables d'interés seleccionades per aquest estudi han estat edat, sexe, dolor de pit, pressió sanguínea, colesterol, sucre a la sang, freqüència cardíaca màxima, angina induïda i, necessàriament, target (presència de dolència cardíaca).

Aquestes variables han estat escollides per adequació dels coneixements en el camp dels membres del grup i per la seva importància a priori com a indicatives de l'estat de salut.

## 3. Neteja de les dades.

### 3.1. Valors nuls

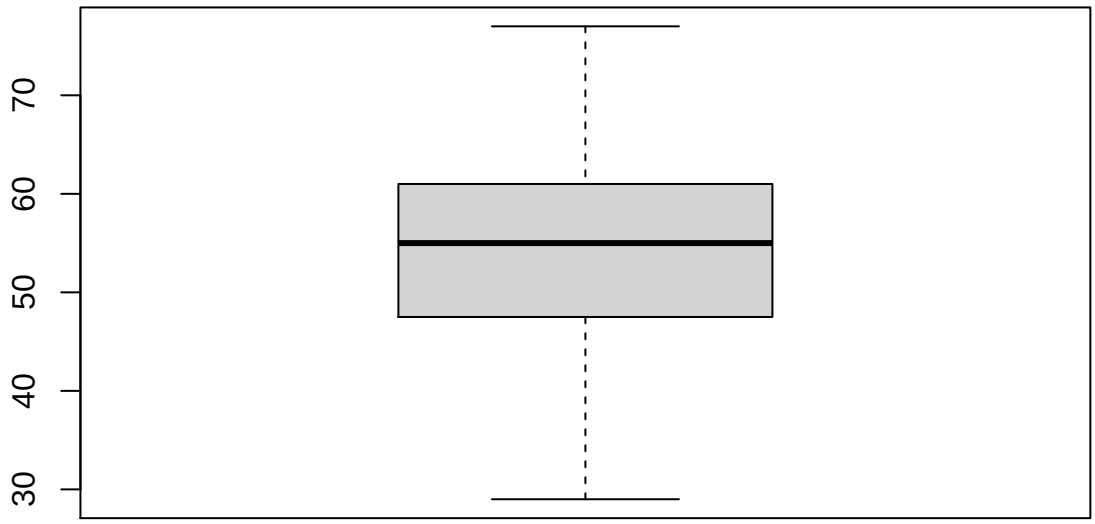
```
sapply(data2, function(x) sum(is.na(x)))
```

```
##           age           sex     chestPain bloodPressure   cholesterol
##             0             0             0             0             0
##    bloodSugar  maxHeartRate     indAngina        target
##             0             0             0             0
```

Com podem veure, no hi ha valors nuls. Tanmateix cal clarificar que si ens trobessim algun cas la millor estratègia hagués estat la seva imputació probabilística pel mètode del veí més proper. Aquest tipus d'estratègia seria l'idonea pel tipus de dades que manegem i per l'escassetat d'observacions.

### 3.2. Valors atípics.

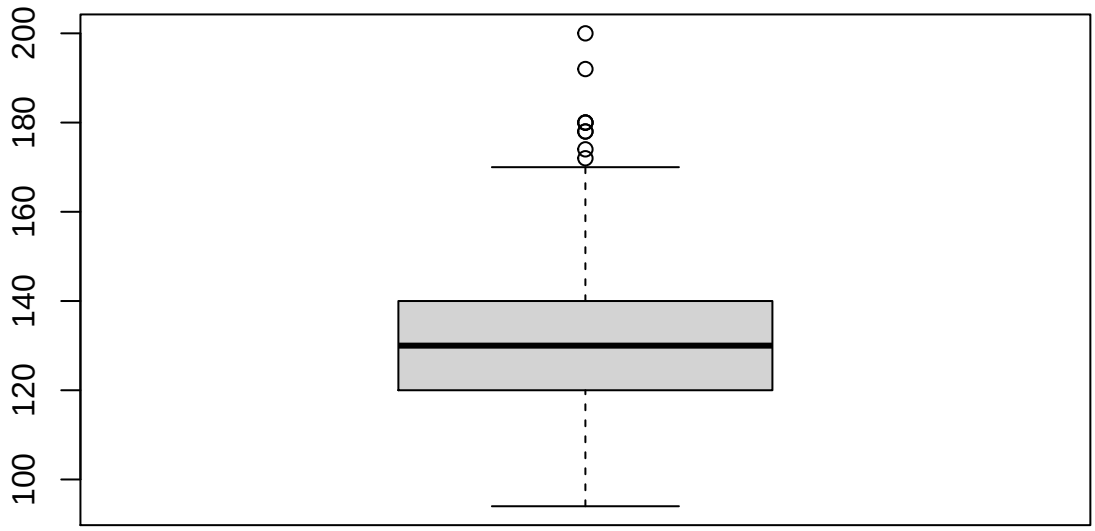
```
boxplot(data2$age)
```




```
bpage <- quantile(data2$age, 0.75) - quantile(data2$age, 0.25)
sum(data2$age > median(data2$age)+3*bpage)
```

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

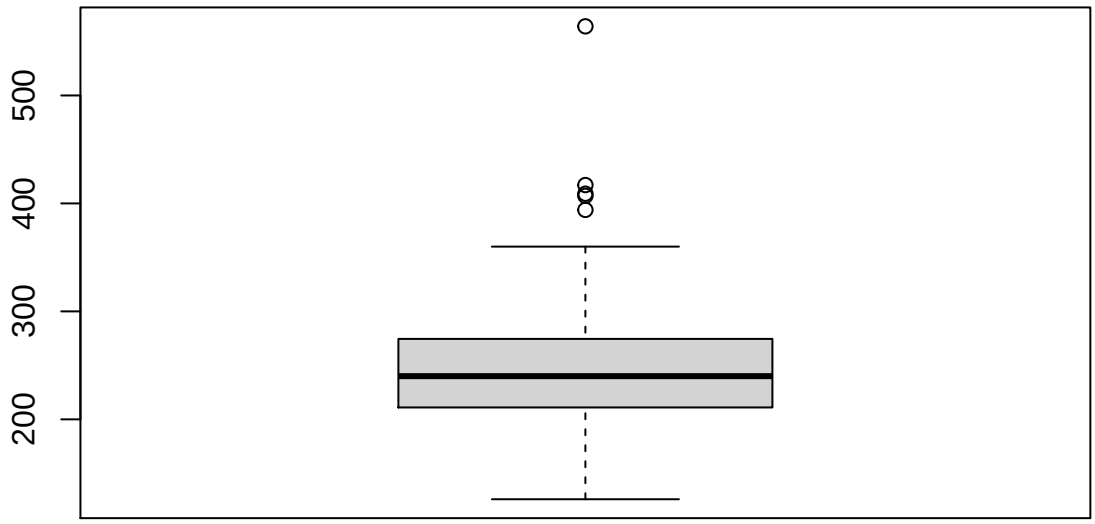
```
boxplot(data2$bloodPressure)
```




```
bprq <- quantile(data2$bloodPressure, 0.75) - quantile(data2$bloodPressure, 0.25)
sum(data2$bloodPressure > median(data2$bloodPressure)+3*bprq)
```

```
## [1] 2
```

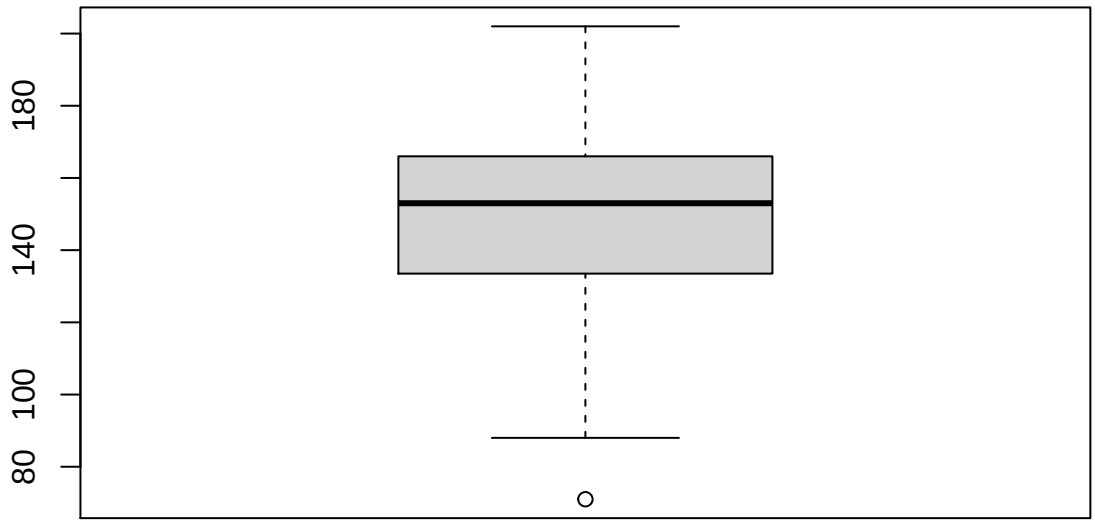
```
boxplot(data2$cholesterol)
```




```
chorq <- quantile(data2$cholesterol, 0.75) - quantile(data2$cholesterol, 0.25)
sum(data2$cholesterol > median(data2$cholesterol)+3*chorq)
```

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

```
boxplot(data2$maxHeartRate)
```




```
mhrrq <- quantile(data2$maxHeartRate, 0.75) - quantile(data2$maxHeartRate, 0.25)
sum(data2$maxHeartRate < median(data2$maxHeartRate)-3*mhrrq)
```

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

Com podem observar als boxplots i als recomptes de valors atípics veiem que existeixen valors anòmals per tres de les variables quantitatives i valors extremadament anòmals per dues d'elles. Tanmateix una investigació del comportament d'aquestes variables pel cas general podem afirmar que, si bé són valors llunyans del rang interquartílic, són valors perfectament possibles i que no és probable que es deguin a errors de mesura. Per tant, decidim mantenir-los en la nostra anàlisi.

```
#### Exportació #####
heart <- data2
write.csv(heart,"heart_clean.csv")
```

Un cop realitzada la neteja i selecció de les dades, procedim a exportar el dataset definitiu.

## 4. Anàlisi de les dades.

### 4.1. Planificació de l'anàlisi

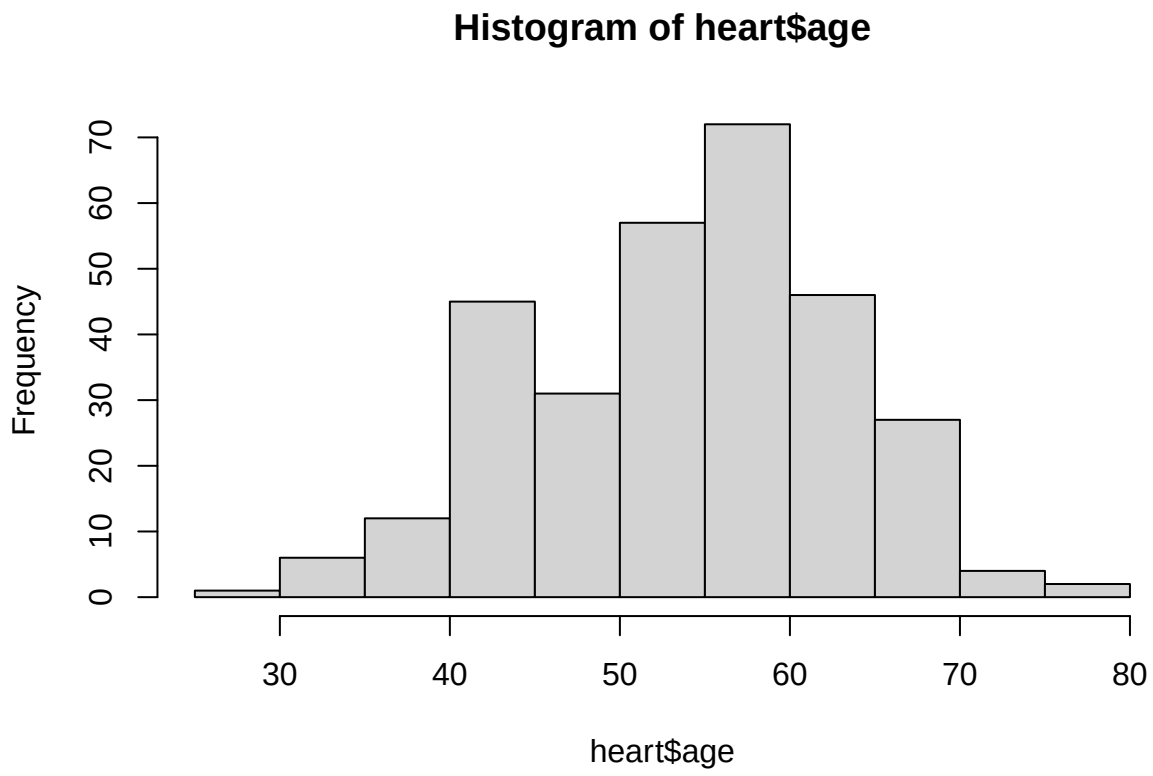
L'anàlisi que realitzarem de les dades es dividirà en tres subapartats. Primerament analitzarem la correlació existent entre les variables quantitatives. Aquesta anàlisi la realitzarem per descobrir si existeix multicolinealitat entre les dades i conèixer millor la relació interna entre aquestes variables. Seguidament, realitzarem tests de chi-quadrat per evaluar la significació de la relació entre les variables categòriques i la variable dependent (target). Finalment, realitzarem una anàlisi de possibles regressions logístiques i evaluarem quin és el millor model explicatiu i predictiu per les dades. Tanmateix, prèviament analitzarem la normalitat i homogeneitat de les variables quantitatives.

### 4.2. Normalitat i Homogeneïtat

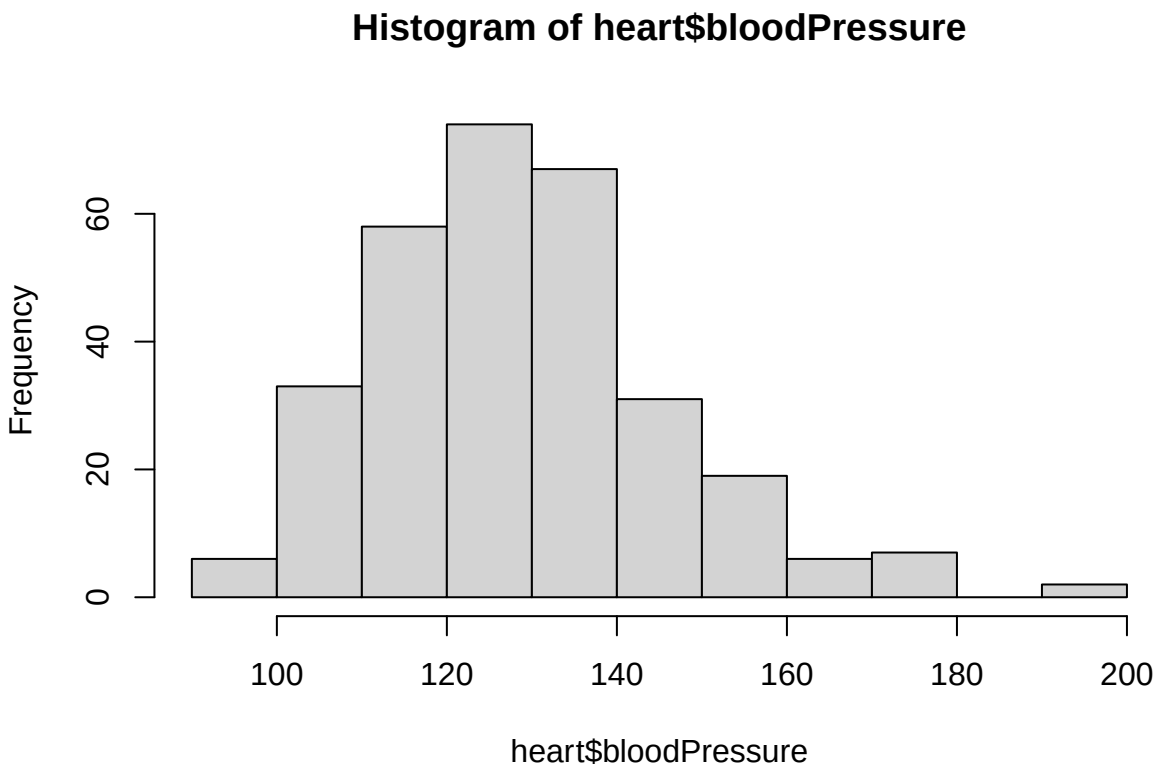
```
# Shapiro-Wilks (Normalitat)
normtest <- data.frame(matrix(ncol = 2, nrow = 4))
colnames(normtest) <- c("Nom", "Pvalor")
normtest[1,1]<-"Edat"
normtest[2,1]<-"Pressió"
normtest[3,1]<-"Colesterol"
normtest[4,1]<-"Freqüència"
normtest[1,2]<-shapiro.test(heart$age)$p.value # No hi ha normalitat
normtest[2,2]<-shapiro.test(heart$bloodPressure)$p.value  # No hi ha normalitat
normtest[3,2]<-shapiro.test(heart$cholesterol)$p.value  # No hi ha normalitat
normtest[4,2]<-shapiro.test(heart$maxHeartRate)$p.value  # No hi ha normalitat
normtest
```

```
##          Nom       Pvalor
## 1       Edat 5.798359e-03
## 2    Pressió 1.458097e-06
## 3 Colesterol 5.364848e-09
## 4 Freqüència 6.620819e-05
```

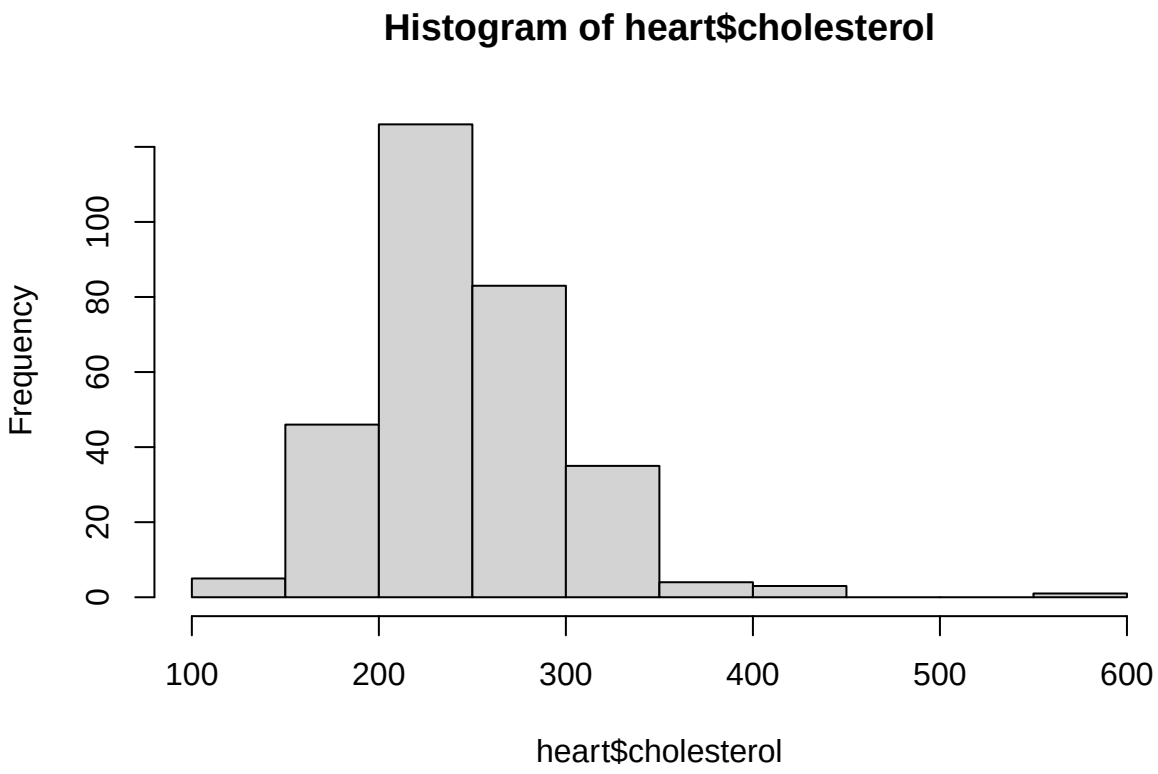
```
hist(heart$age)
```



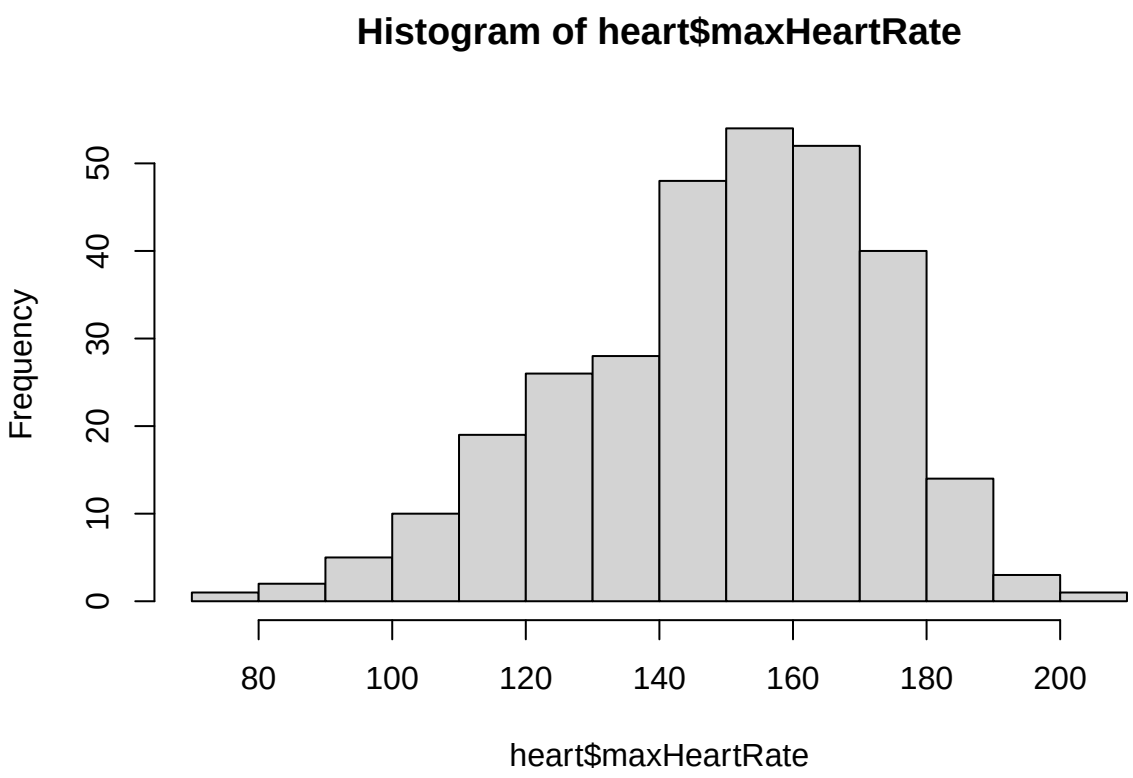

```
hist(heart$bloodPressure)
```




```
hist(heart$cholesterol)
```




```
hist(heart$maxHeartRate)
```




Com podem veure pels resultats dels tests Shapiro-Wilks i corroborar per la forma dels histogrames, ninguna de les variables quantitatives presenta normalitat per les seves distribucions. El que comportarà aquesta situació és la utilització d'anàlisis no paramètrics.

```
hov(heart$age ~ heart$target) # No hi ha homogeneïtat
```

```
##
##  hov: Brown-Forsyth
##
## data:  heart$age
## F = 7.9854, df:heart$target = 1, df:Residuals = 301, p-value = 0.005031
## alternative hypothesis: variances are not identical
```

```
hov(heart$bloodPressure ~ heart$target) # Hi ha homogeneïtat
```

```
##
##  hov: Brown-Forsyth
##
## data:  heart$bloodPressure
## F = 1.857, df:heart$target = 1, df:Residuals = 301, p-value = 0.174
## alternative hypothesis: variances are not identical
```

```
hov(heart$cholesterol ~ heart$target) # Hi ha homogeneïtat
```

```
##
##  hov: Brown-Forsyth
##
## data:  heart$cholesterol
## F = 0.10146, df:heart$target = 1, df:Residuals = 301, p-value = 0.7503
```

```
## alternative hypothesis: variances are not identical
```

```
hov(heart$maxHeartRate ~ heart$target) # No hi ha homogeneïtat
```

```
##
##  hov: Brown-Forsyth
##
## data:  heart$maxHeartRate
## F = 5.2467, df:heart$target = 1, df:Residuals = 301, p-value = 0.02268
## alternative hypothesis: variances are not identical
```

Com veiem pel test Brown-Forsyth, adequat per distribucions no normals, només podem afirmar amb significació que existeix homogeneïtat per les variables pressió sanguínea i colesterol. Aquests resultats tindrà relevància en les decisions referents a l'anàlisi de correlacions.