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. I a més, intentar crear un model per predir-les.

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

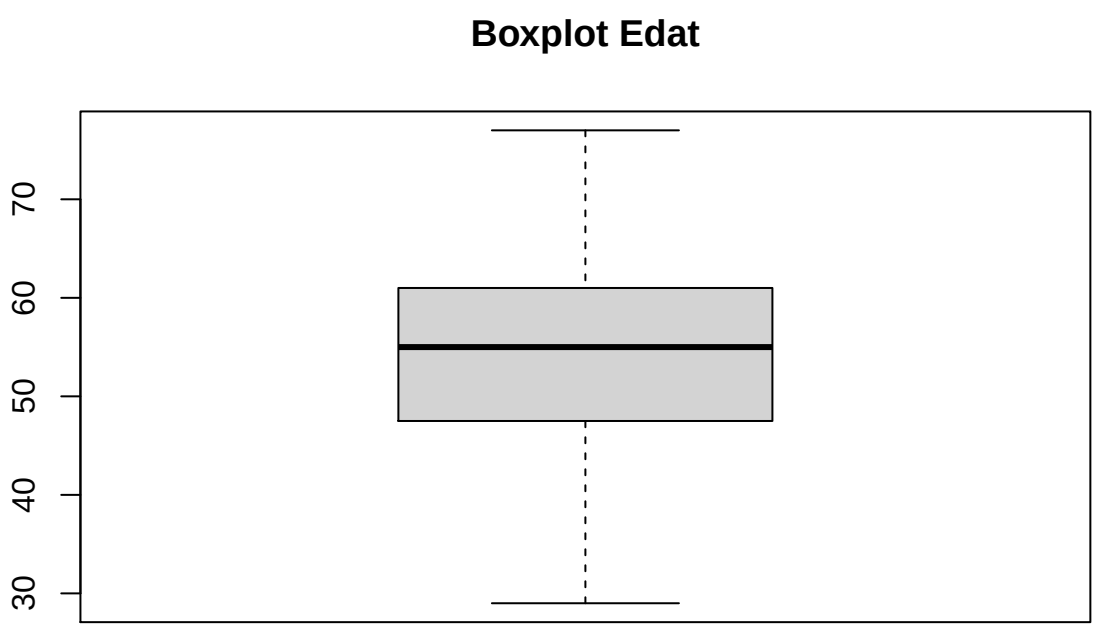
```
sapply(data2, function(x) sum(is.na(x)))
```

```
##           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.

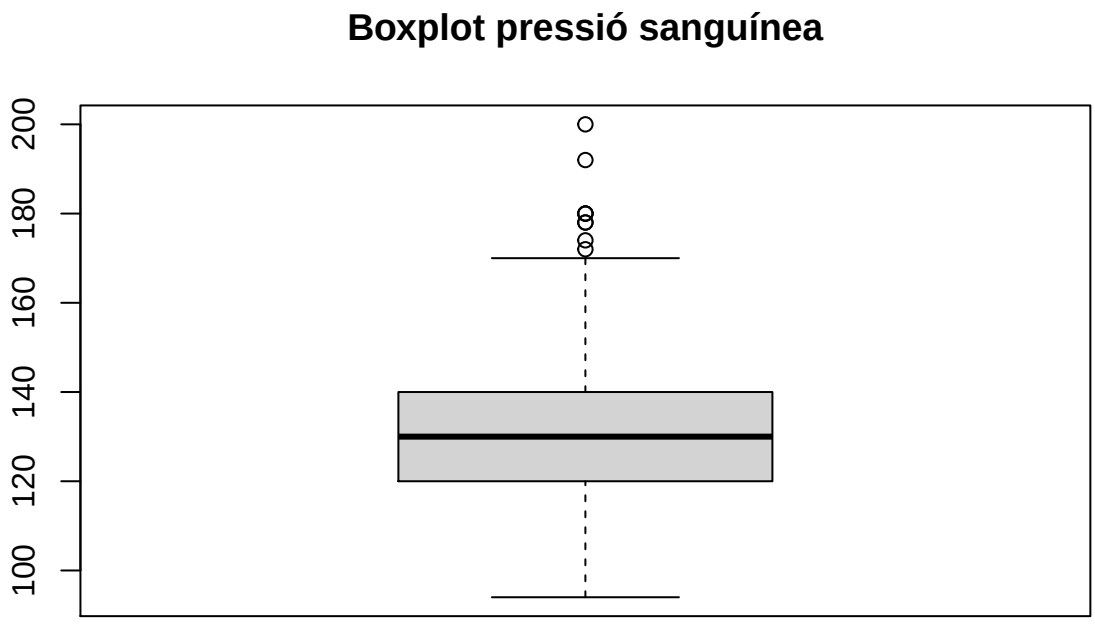
```
boxplot(data2$age, main = "Boxplot Edat")
```




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

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

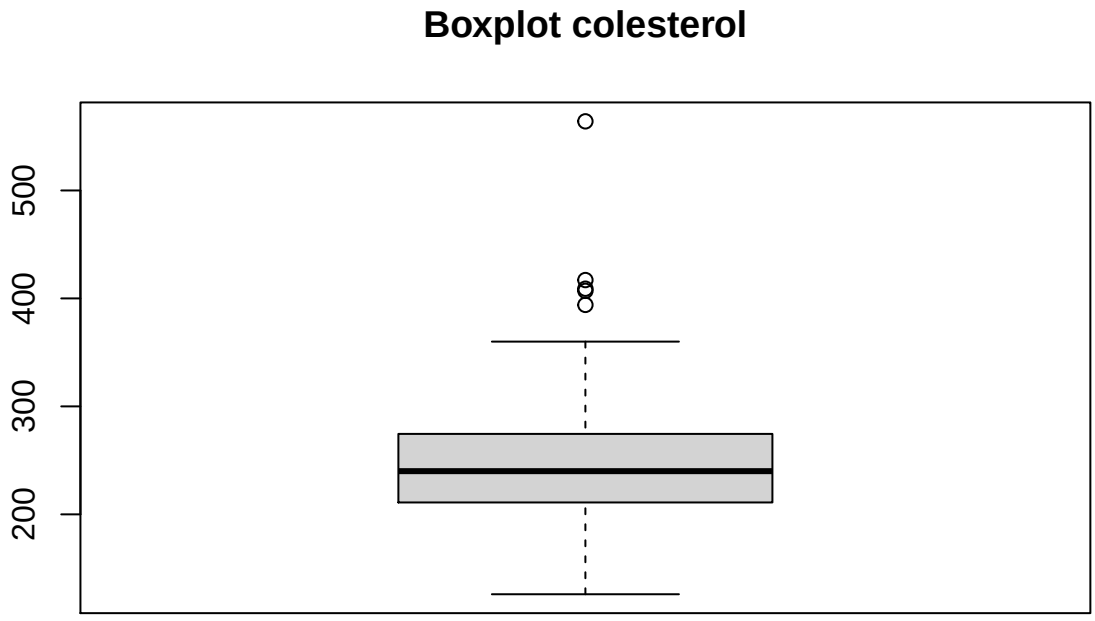
```
boxplot(data2$bloodPressure, main = "Boxplot pressió sanguínea")
```




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

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

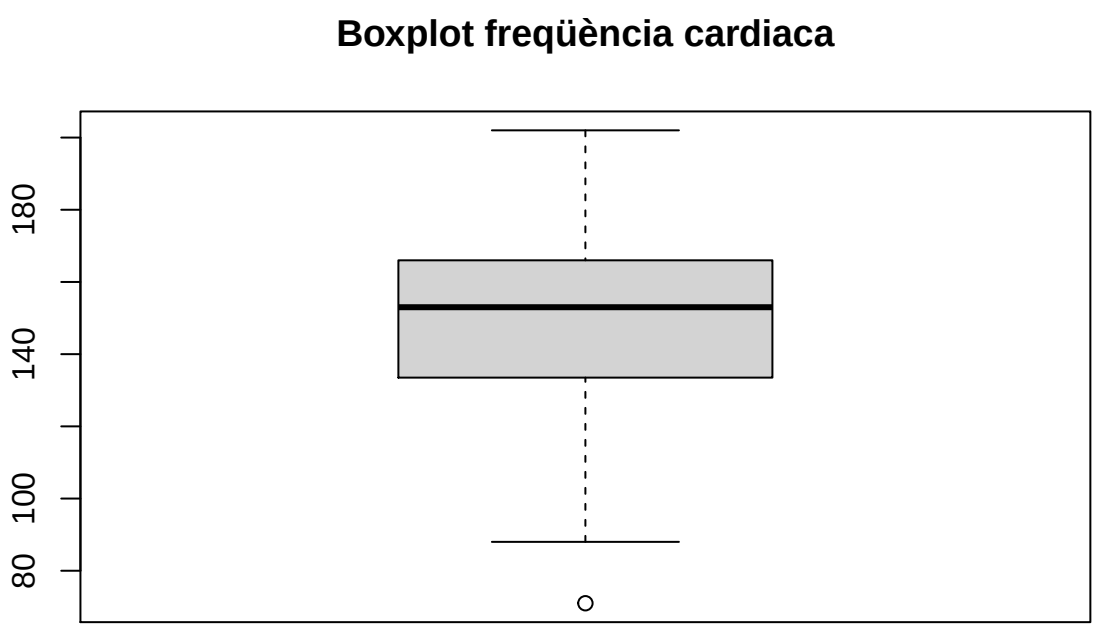
```
boxplot(data2$cholesterol, main = "Boxplot colesterol")
```




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

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

```
boxplot(data2$maxHeartRate, main = "Boxplot freqüència cardiaca")
```




```
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.

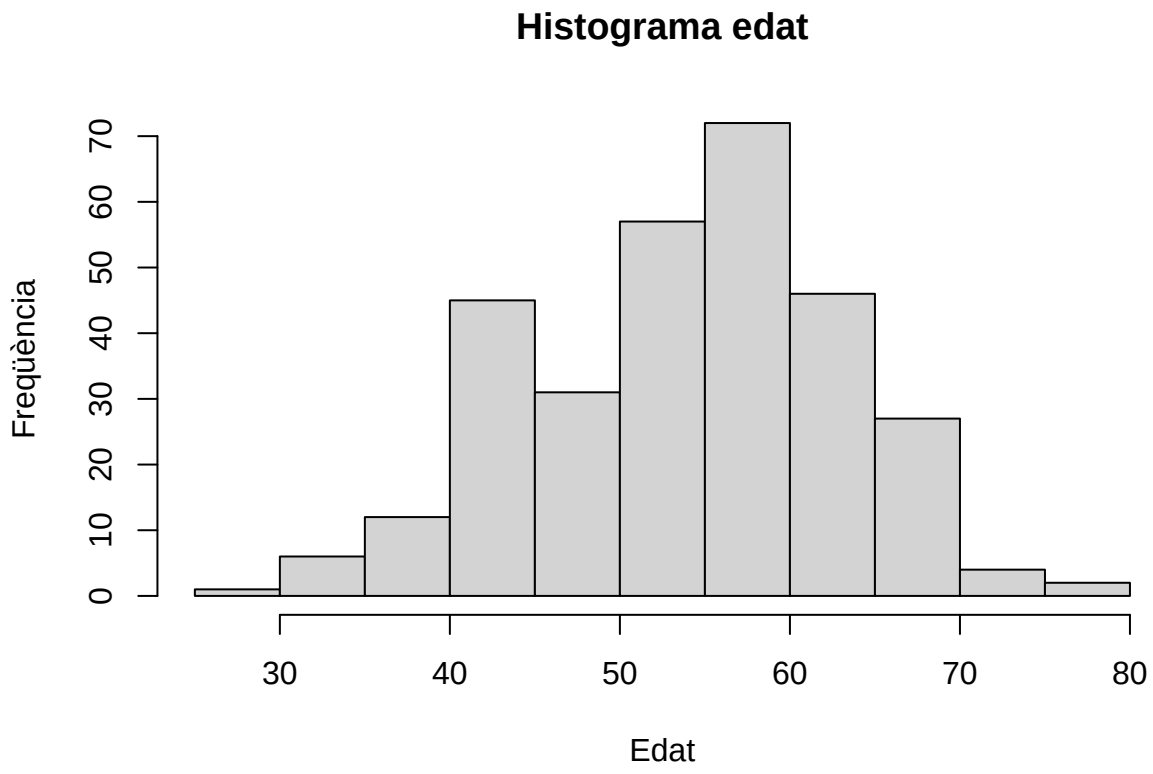
Abans de res, prèviament analitzarem la normalitat i homogeneitat de les variables quantitatives.

4.2. Normalitat i Homogeneïtat

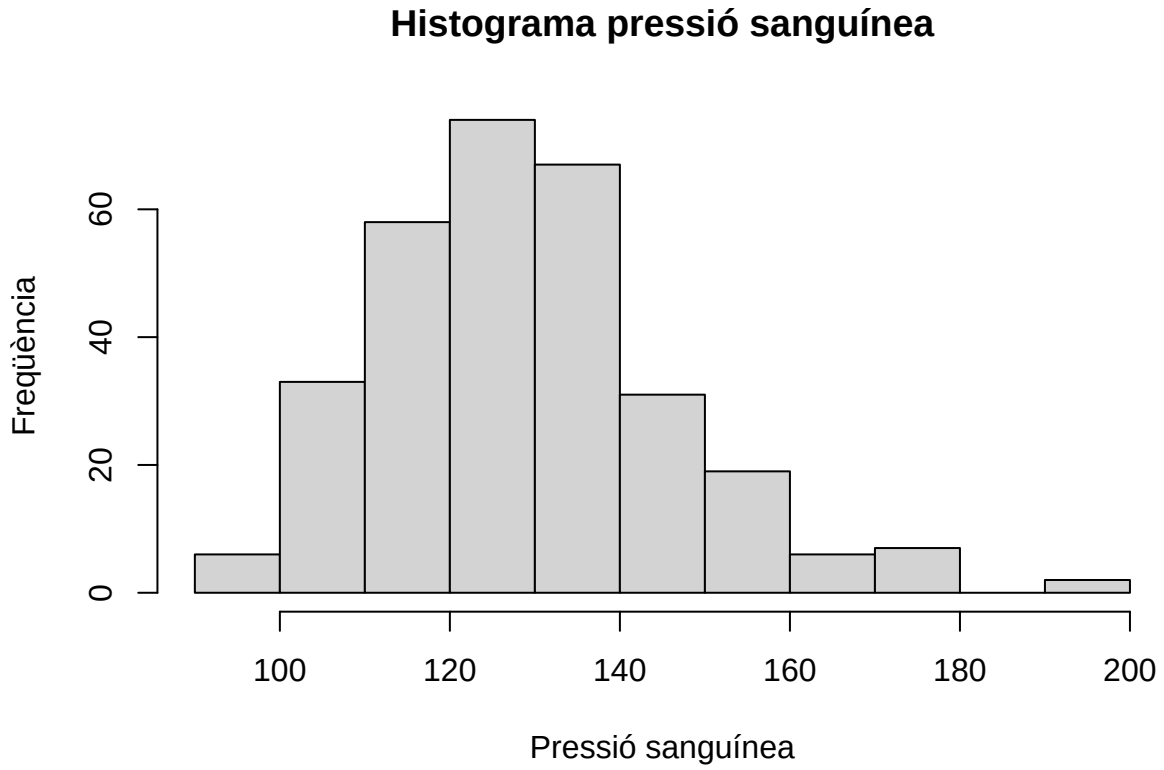
```
# Shapiro-Wilks (Normalitat)
normtest <- data.frame(matrix(ncol = 2, nrow = 4))
colnames(normtest) <- c("Nom", "Pvalor")
normtest[1,1]<-"Edat"
normtest[2,1]<-"Pressió"
normtest[3,1]<-"Colesterol"
normtest[4,1]<-"Freqüència"
normtest[1,2]<-shapiro.test(heart$age)$p.value # No hi ha normalitat
normtest[2,2]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

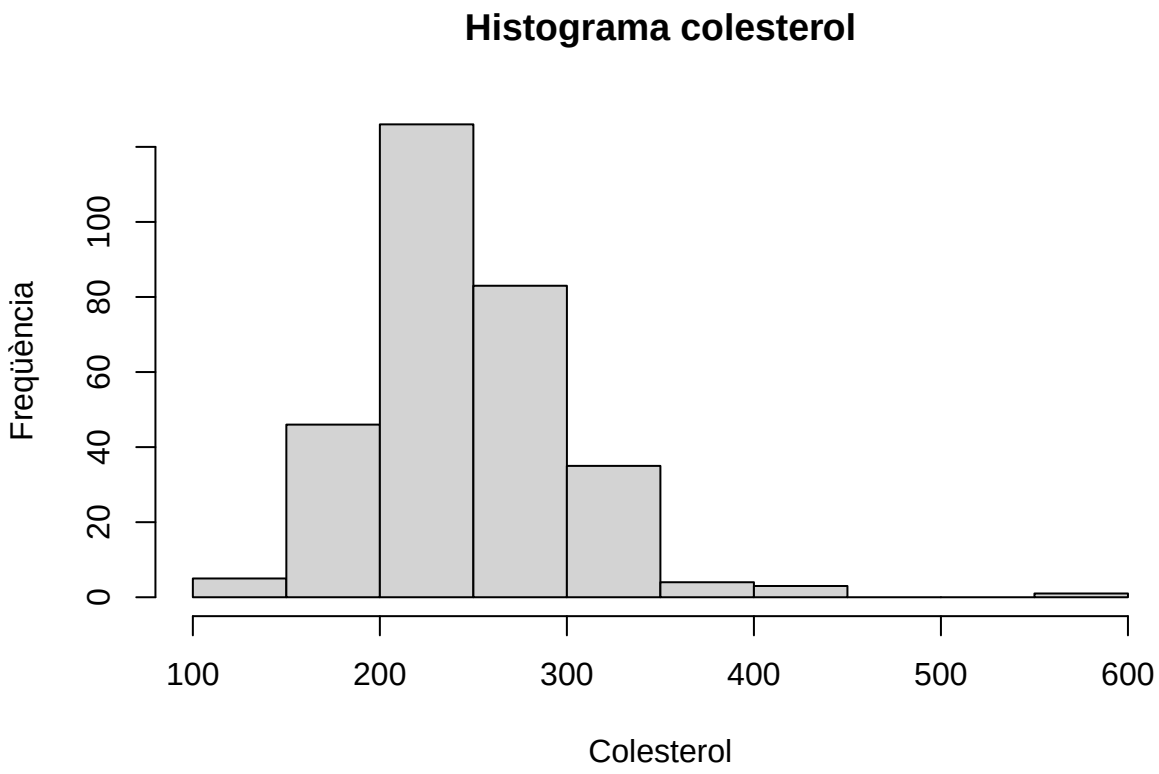
```
hist(heart$age, main = "Histograma edat", ylab = "Freqüència", xlab = "Edat")
```



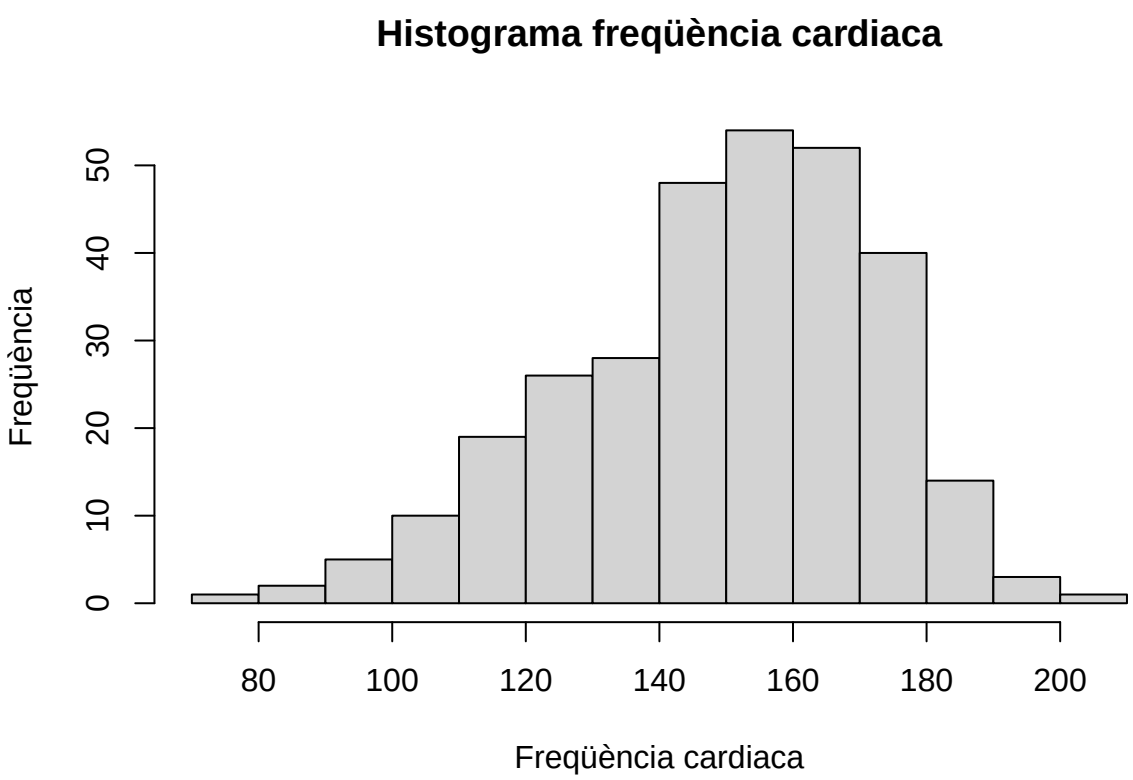
```r
hist(heart$bloodPressure, main = "Histograma pressió sanguínea", ylab = "Freqüència", xlab = "Pressió sa
```




```r
hist(heart$cholesterol, main = "Histograma colesterol", ylab = "Freqüència", xlab = "Colesterol")
```




```
hist(heart$maxHeartRate, main = "Histograma freqüència cardiaca", ylab = "Freqüència", xlab = "Freqüènc
```



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 ja que presenten un p-valor major al 0,05 (prenem confiança del 95%). Aquests resultats tindrà relevància en les decisions referents a l'anàlisi de correlacions.

4.3 Tres mètodes d'anàlisi diferents.

4.3.1 Correlació entre les variables numèriques. Per tal de comprovar la possible existència de correlació entre alguna de les variables numèriques, anem a fer un test de correlacions per parelles 2 a 2.

```
pvalorcor<-c(1:6)
mcor <- cor(heart[,c(1,4,5,7)], method = "kendall")
c<-1
for(i in c(1,4,5,7)){
  for(j in c(1,4,5,7)){

    if(i!=j & i<j ){ pvalorcor[c] <- cor.test(heart[,i],heart[,j], method = "kendall")$p.value
    c <- c+1 }
  }
}
Varcor <- c("Age-bloodPressure","Age-Cholesterol","Age-HeartRate","bloodPressure-cholesterol","bloodPres
cordf <- data.frame(Varcor, pvalorcor)
cordf
```

```
##                      Varcor    pvalorcor
## 1         Age-bloodPressure 5.367436e-07
## 2           Age-Cholesterol 5.683979e-04
## 3             Age-HeartRate 1.096568e-12
## 4 bloodPressure-cholesterol 2.912858e-02
## 5   bloodPressure-heartRate 4.853321e-01
## 6     cholesterol-heartRate 4.186589e-01
```

Observem que en tots casos el p-valor és menor a 0,05 (95% de confiança). Com en tots cassos és inferior a 0.05, tenim prou evidència per dir que la correlació entre les parelles variables és estadísticament significativa.

Vegem però, que les parelles de bloodPressure-heartRate i cholesterol-heartRate són les que a priori presentaran una correlació menor, ja que el seu p-valor és més pròxim al 0.05.

4.3.2 Test Chi-quadrat Com hem comentat anem a fer tests de chi-quadrat per evaluar la significació de la relació entre les variables categòriques i la variable dependent (target).

```
pvalor_chi_quadrat <- c(1:4)
c <- 1
for(i in c(2,3,6,8)){
  pvalor_chi_quadrat[c] <- chisq.test(heart$target,heart[,i])$p.value
  c <- c+1
}
```

```r
Variables <- c("Target-Sex","Target-ChestPain","Target-Bloodsugar","Target-IndAngina")
chidf <- data.frame(Variables, pvalor_chi_quadrat )
chidf
```

```
##           Variables pvalor_chi_quadrat
## 1        Target-Sex       1.876778e-06
## 2  Target-ChestPain       1.334304e-17
## 3 Target-Bloodsugar       7.444281e-01
## 4  Target-IndAngina       7.454409e-14
```

Excepte en el cas del Target-Bloodsugar, els p-valors són inferiors a 0,05, per la qual cosa rebutgem la hipòtesi nul · la d'independència entre les variables 2 a 2. Per tant, els nostres factors es correlacionen ja que són depenents.

Per tant, el Bloodsugar és l'única variable significativament independent del Target.

4.3.3 Regressió Logística Anem a crear un primer model amb totes les variables:

```r
model1 <- glm(target~ ., data = heart, family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = target ~ ., family = "binomial", data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6925  -0.5990   0.2271   0.6682   2.7116
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.803522   2.149063   0.839 0.401350
## age                        -0.025870   0.020581  -1.257 0.208766
## sexMale                    -1.987022   0.392128  -5.067 4.04e-07 ***
## chestPainAtypical Angina    1.850498   0.390392   4.740 2.14e-06 ***
## chestPainNon-Anginal Pain   1.919092   0.586041   3.275 0.001058 **
## chestPainTypical Angina     1.869227   0.491151   3.806 0.000141 ***
## bloodPressure              -0.021150   0.009434  -2.242 0.024975 *
## cholesterol                -0.006157   0.003179  -1.937 0.052760 .
## bloodSugar> 120 mg/dl       0.002543   0.439905   0.006 0.995388
## maxHeartRate                0.032196   0.009060   3.554 0.000380 ***
## indAnginaYes               -0.865614   0.368357  -2.350 0.018777 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 259.63  on 292  degrees of freedom
## AIC: 281.63
##
## Number of Fisher Scoring iterations: 5
```

```r
#predict(object=model1, newdata=heart_test_vars)
```

Com hem vist en el test Chi-quadrat el Bloodsugar és l'única variable significativament independent del

Target, per tant la llevam del model.

A més, també llevam l'edat perquè ens indica no significació el model. Té sentit amb el nostre anàlisi de correlació doncs l'edat té relació amb la resta de variables quantitatives i afagir-la seria redundant, en altres paraules, no importa l'edat sinó lo 'malament' que estàs.

```
model2 <- glm(target~.-bloodSugar - age, data = heart, family = "binomial")
summary(model2)
```

```
##
## Call:
## glm(formula = target ~ . - bloodSugar - age, family = "binomial",
##     data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7338  -0.5976   0.2281   0.6800   2.6837
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.283512   1.759184   0.161 0.871967
## sexMale                    -1.936688   0.388203  -4.989 6.07e-07 ***
## chestPainAtypical Angina    1.837602   0.383786   4.788 1.68e-06 ***
## chestPainNon-Anginal Pain   1.885801   0.583206   3.234 0.001223 **
## chestPainTypical Angina     1.858594   0.487239   3.815 0.000136 ***
## bloodPressure              -0.024342   0.009051  -2.690 0.007155 **
## cholesterol                -0.006823   0.003154  -2.163 0.030504 *
## maxHeartRate                0.036468   0.008438   4.322 1.55e-05 ***
## indAnginaYes               -0.839136   0.364816  -2.300 0.021439 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 261.23  on 294  degrees of freedom
## AIC: 279.23
##
## Number of Fisher Scoring iterations: 5
```

Obtenim així que el model 2 és millor, ja que l'AIC del model 1 és superior al del model 2. 281,63 enfront de 279,23.

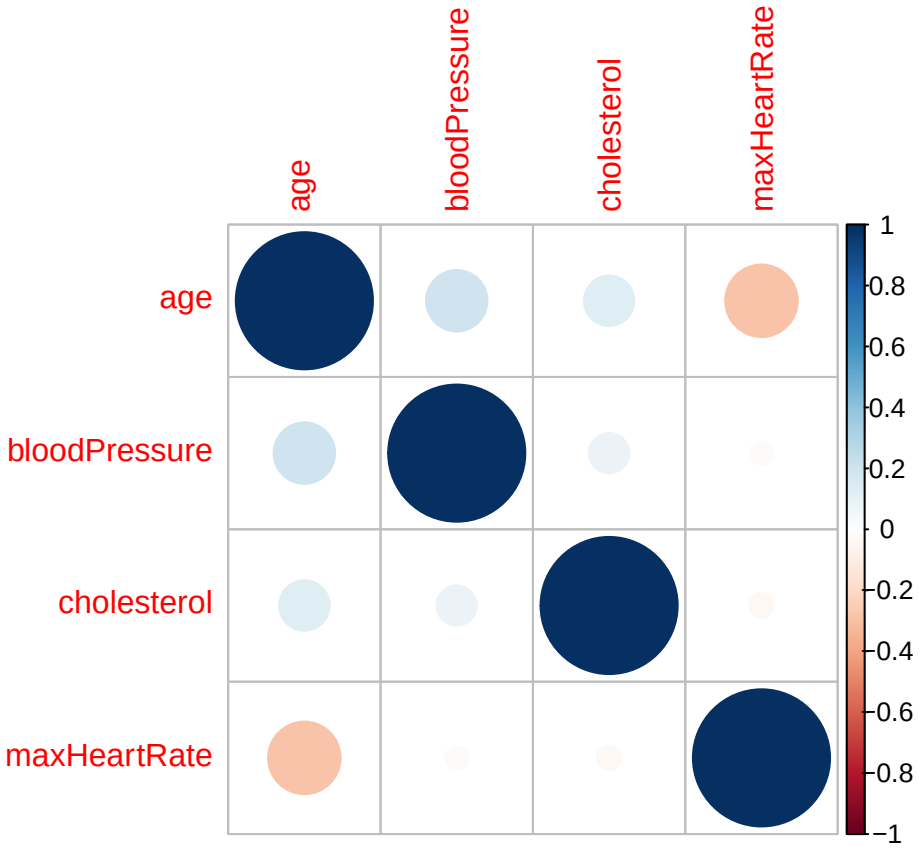
5. Representació dels resultats a partir de taules i gràfiques

De les tres proves estadístiques realitades en l'apartar anterior anem a representar els resultats a partir de taules i gràfiques.

5.1 Representació de la correlació entre les variables numèriques.

Fem la matriu de correlació entre les variables numèriques del dataset.

```
corrplot(mcor)
```



Efectivament, com ja haviem vist amb el test de correlacions, les parelles de bloodPressure-heartRate i cholesterol-heartRate són les que tenen una menor correlació entre elles.

La parella amb més correlació és la age-maxHeartRate.

5.2 Taules de contingència usades en el Test Chi-quadrat

Anem a veure les taules de contingència usades en el 4 test de Chi-quadrat fets.

```
# Target-Sex:
Target_Sex = table(heart$sex,heart$target)
print(Target_Sex)
```

```
##
##           NO YES
##   Female  24  72
##   Male   114  93
```

```
# Target-ChestPain:
Target_ChestPain = table(heart$chestPain,heart$target)
print(Target_ChestPain)
```

```
##
##                     NO YES
##   Asymptomatic     104  39
##   Atypical Angina   18  69
##   Non-Anginal Pain   7  16
##   Typical Angina     9  41
```

```r
# Target-Bloodsugar:
Target_Bloodsugar = table(heart$bloodSugar,heart$target)
print(Target_Bloodsugar)
```

```
##
##                 NO YES
##   < 120 mg/dl 116 142
##   > 120 mg/dl  22  23
```

```r
# Target-IndAngina:
Target_IndAngina = table(heart$indAngina,heart$target)
print(Target_IndAngina)
```

```
##
##        NO YES
##   No   62 142
##   Yes  76  23
```

Comprovem que la taula de contingència entre Target-Bloodsugar (la tercera) és la que ménys relació s'observa entre les dues variables, com ja havíem vist amb el test Chi-quadrat.

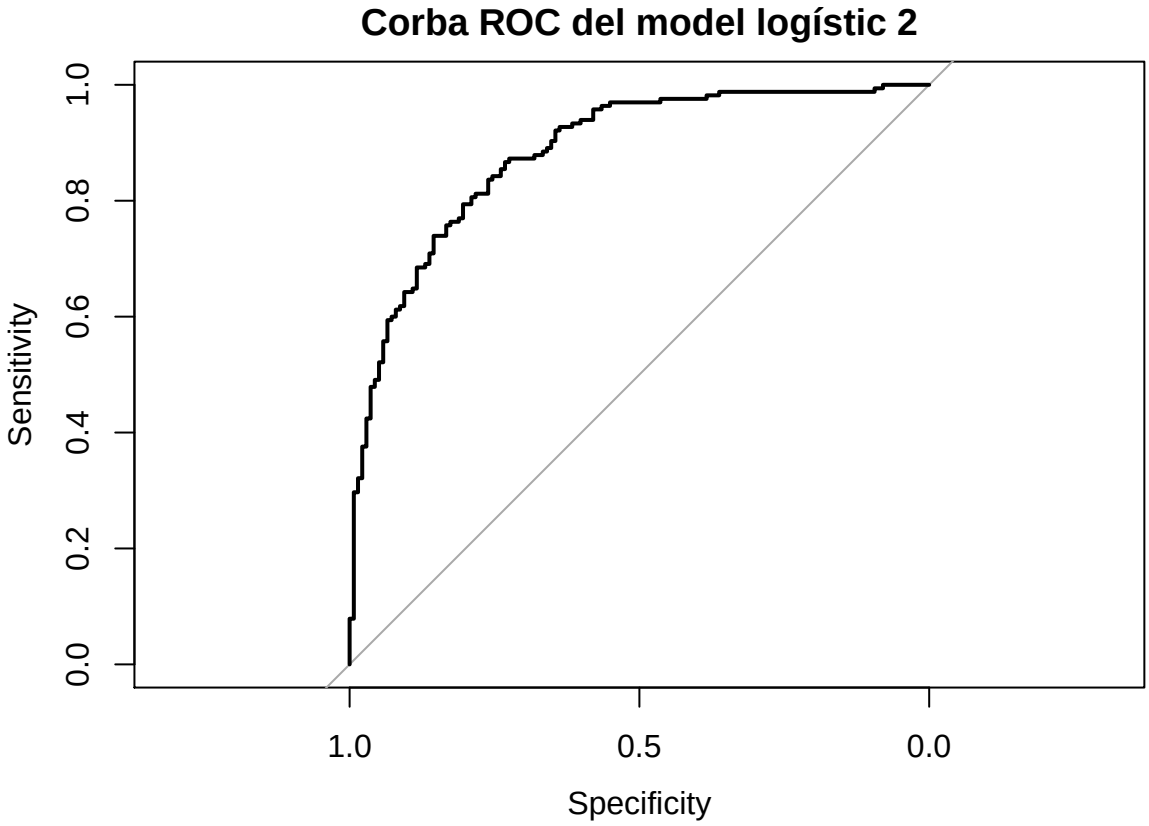
5.3 Regressió Logística

```r
library(pROC)
prob = predict(model2, heart, type="response")
roc_model2 = roc(heart$target, prob, data=heart)
```

```
## Setting levels: control = NO, case = YES
```

```
## Setting direction: controls < cases
```

```r
plot (roc_model2, main= 'Corba ROC del model logístic 2')
```




```
roc_model2
```

```
##
## Call:
## roc.default(response = heart$target, predictor = prob, data = heart)
##
## Data: prob in 138 controls (heart$target NO) < 165 cases (heart$target YES).
## Area under the curve: 0.8824
```

Vegem que l'àrea sota la corba ROC és del 0,8824, bastant propera a 1 que indicaria un model perfecte. Per tant, hem obtingut un model força bo.

6. Resolució del problema

En aquest darrer apartat, anem a partir dels resultats obtinguts, a veure les principals conclusions i a veure com aquestes ens han permès respondre al problema de veure quins factors tenen influència directa sobre l'aparició de malalties coronàries i a predir-les.

Hem vist com es relacionen entre elles les variables numèriques del dataset. Hem vist que bloodPressure-heartRate i cholesterol-heartRate són les que tenen una menor correlació entre elles. A més, cap parella presenta una correlació molt alta.

Llavors, hem observat que les varibales qualitatives: sexe, el ChestPain i la indAngina són dependents amb el target (presència o no de dolència cardíaca), és a dir, estan directament relacionats.

Mentre que el Bloodsugar és independent del target, i l'hem treta del model. També hem tret l'edat ja que el model 1 ens ha indicat no significació amb el target.

Finalment, el model de regressió logística traient les variables menys relacionades amb el target, hem vist

amb la corba ROC, que és un bon model que ens pot ajudar força bé a predir el target, la presència o no de dolència cardíaca.

7. Codi

El codi complet amb R està compartit al Github.

Taula de contribucions al treball:

Contribucions	**Signa**
Investigació prèvia	GMF, AVV
Redacció de les respostes	GMF, AVV
Desenvolupament codi	GMF, AVV