Heart Attack Analysis & Prediction Dataset

Pràctica 2 - Tipologia i cicle de vida de les dades

Francesc Valls i Pol Codinachs

Càrrega de la base de dades que s'utilitza per aquesta pràctica

```
taula_heart = read.csv("heart.csv")
```

Verifiquem l'estructura del joc de dades principal on veiem que conté un total de: *303 registres* 14 variables

```
str(taula_heart)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

1. Descripció del dataset

Aquestes són les variables que conté el dataset:

- Age: Variable numèrica, indica l'edat del pacient.
- Sex: Variable booleana, indica el sexe del pacient (0 = Dones, 1=Homes)
- cp: Variable numèrica, indica el tipus de dolor al pit (Chest pain):
  - Valor 1: angina típica
  - Valor 2: angina atípica
  - Valor 3: sense angina
  - Valor 4: assimptomatic
- trtbps: Variable numèrica, indica la pressió sanguínia en repòs, en mm de Hg
- chol: Variable numèrica, indica el colesterol en mg/dl obtingut a través del sensor BMI
- fbs: Variable booleana, indica si la quantitat de sucre a la sang és > 120 mg/dl:
  - Valor 0: Fals
  - Valor 1: Verdader
- rest_ecg: Variable booleana, indica els resultats electrocardiogràfics en repòs:
  - Valor 0: Normal
  - Valor 1: Anomalíes en la corba ST-T
- thalachh: Variable numèrica, indica la freqüència cardíaca màxima registrada
- exng: Variable booleana: indica angina induida per l'exercici
  - Valor 0: No
  - Valor 1: Si
- oldpeak: Variable numèrica, indica el "peak" anterior
- slp: Variable numèrica,
- caa: Els vasos que retornen la sang al cor
- thall: No explica el significat de la variable en el kaggle
- output: Variable dicotòmica
  - valor 0: No ha patit un atac de cor
  - valor 1: Si ha patit un atac de cor

Amb la descripció del dataset podem observar quines variables poden ser rellevants per realitzar els anàlisis pertinents per veure si tenen sentit analitzar-les per predir les persones que podrien patir un atac de cor.

Declarem les variables que són factors:

```
taula_heart$sex <- as.factor(taula_heart$sex)
taula_heart$cp <- as.factor(taula_heart$cp)
taula_heart$fbs <- as.factor(taula_heart$fbs)
taula_heart$exng <- as.factor(taula_heart$exng)
taula_heart$output <- as.factor(taula_heart$output)

summary(taula_heart)
```

```
##       age        sex     cp          trtbps           chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
##     restecg          thalachh     exng       oldpeak          slp
##  Min.   :0.0000   Min.   : 71.0   0:204   Min.   :0.00   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:133.5   1: 99   1st Qu.:0.00   1st Qu.:1.000
##  Median :1.0000   Median :153.0           Median :0.80   Median :1.000
##  Mean   :0.5281   Mean   :149.6           Mean   :1.04   Mean   :1.399
```

```
##  3rd Qu.:1.0000   3rd Qu.:166.0          3rd Qu.:1.60   3rd Qu.:2.000
##  Max.   :2.0000   Max.   :202.0          Max.   :6.20   Max.   :2.000
##       caa             thall       output
##  Min.   :0.0000   Min.   :0.000   0:138
##  1st Qu.:0.0000   1st Qu.:2.000   1:165
##  Median :0.0000   Median :2.000
##  Mean   :0.7294   Mean   :2.314
##  3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :4.0000   Max.   :3.000
```

2. Integració i selecció

Es crea una nova variable on es discretitza la variable age per trams de 10 anys

```
summary(taula_heart[, "age"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   47.50   55.00   54.37   61.00   77.00
```

```
taula_heart["AgeDisc"] <- cut(taula_heart$age, breaks = c(20,
    30, 40, 50, 60, 70, 80), labels = c("20-29", "30-39", "40-49",
    "50-59", "60-69", "70-79"))

table(taula_heart$fbs)
```

```
##
##   0   1
## 258  45
```

També es modifiquen els noms de les variables per fer-les més entenedores. * age->Age * sex->Sex * cp->ChestPain * trtbps->BloodPres * chol->Cholesterol * fbs->BloodSugar * restecg->RestECG * thalachh->MaxHR * exng->ExAngina * oldpeak->OldPeak * slp->Slope * caa->Vessels * thall->Thall * output->Output

```
names(taula_heart)[names(taula_heart) == "age"] <- "Age"
names(taula_heart)[names(taula_heart) == "sex"] <- "Sex"
names(taula_heart)[names(taula_heart) == "cp"] <- "ChestPain"
names(taula_heart)[names(taula_heart) == "trtbps"] <- "BloodPres"
names(taula_heart)[names(taula_heart) == "chol"] <- "Cholesterol"
names(taula_heart)[names(taula_heart) == "fbs"] <- "BloodSugar"
names(taula_heart)[names(taula_heart) == "restecg"] <- "RestECG"
names(taula_heart)[names(taula_heart) == "thalachh"] <- "MaxHR"
names(taula_heart)[names(taula_heart) == "exng"] <- "ExAngina"
names(taula_heart)[names(taula_heart) == "oldpeak"] <- "OldPeak"
names(taula_heart)[names(taula_heart) == "slp"] <- "Slope"
names(taula_heart)[names(taula_heart) == "caa"] <- "Vessels"
names(taula_heart)[names(taula_heart) == "thall"] <- "Thall"
names(taula_heart)[names(taula_heart) == "output"] <- "HeartAttack"
```

3. Neteja de les dades

Valors buits

A continuació, en l'etapa de neteja de les dades es duen a terme una sèrie de processos que permeten identificar aquells registres incomplets, incorrectes, inexactes o no pertinents del nostre joc de dades per tal d'eliminar-los o bé corregir-los. Amb tot això aconseguim millorar la qualitat de les dades.

```
nas <- sum(is.na(taula_heart))
paste("Valors NA: ", nas)
```

```
## [1] "Valors NA:  0"
```

Zeros

Veiem que en aquest dataset no existeixen valors en nuls pel que no cal tractar-los, en el cas que aquests existissin, els podriem tractar de dues maneres diferents:

*Si són pocs els registres blancs o nuls, eliminar-los ja que l'impacte que tindrien sobre el conjunt total seria mínim.

*En el cas que n'hi hagués un nombre considerable, es podrien emplenar per l'string "Desconegut" aquells atributs del tipus char o per la mitjana de tots els valors de l'atribut aquells que siguin del tipus numèric.

Anem a veure també, d'aquelles variables numèriques (no booleanes ni categòriques) *age*, *trtbps*, *chol*, *thalachh*, quants valors igual a zero tenen.

Observem que cap d'aquestes variables té cap valor igual a zero.

```
colSums(taula_heart[c("Age", "BloodPres", "Cholesterol", "MaxHR")] ==
    0)
```

```
##         Age   BloodPres Cholesterol       MaxHR
##           0           0           0           0
```

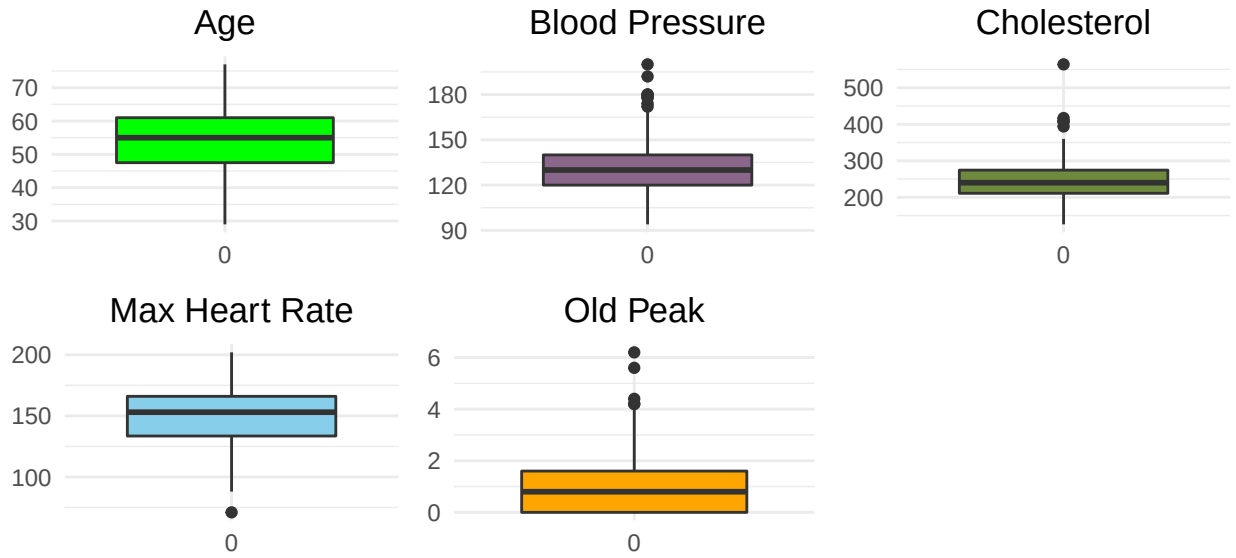
Valors extrems

Anem a veure els valors extrems (outliers) que conté el joc de dades. Per això, seleccionem les variables numèriques i descartem les booleanes i amb l'ajuda dels boxplots visualitzem la distribució de cada variable. Podem observar que el nombre de valors extrems és molt petit en proporció la quantitat de registres totals i al tractar-se de dades fisiològiques, on cada pacient és un cas totalment diferent, poden considerar-se normals.

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)

taula_heart_age.bp <- ggplot(data = data.frame(taula_heart$Age),
    aes(x = factor(0), y = taula_heart$Age)) + theme_minimal() +
    geom_boxplot(fill = "Green") + theme(axis.title.x = element_blank(),
    axis.title.y = element_blank()) + ggtitle("Age") + theme(plot.title = element_text(hjust = 0.5))
taula_heart_trtbps.bp <- ggplot(data = data.frame(taula_heart$BloodPres),
    aes(x = factor(0), y = taula_heart$BloodPres)) + theme_minimal() +
    geom_boxplot(fill = "plum4") + theme(axis.title.x = element_blank(),
    axis.title.y = element_blank()) + ggtitle("Blood Pressure") +
    theme(plot.title = element_text(hjust = 0.5))
taula_heart_chol.bp <- ggplot(data = data.frame(taula_heart$Cholesterol),
    aes(x = factor(0), y = taula_heart$Cholesterol)) + theme_minimal() +
    geom_boxplot(fill = "darkolivegreen4") + theme(axis.title.x = element_blank(),
    axis.title.y = element_blank()) + ggtitle("Cholesterol") +
    theme(plot.title = element_text(hjust = 0.5))
taula_heart_thalachh.bp <- ggplot(data = data.frame(taula_heart$MaxHR),
```

```
    aes(x = factor(0), y = taula_heart$MaxHR)) + theme_minimal() +
    geom_boxplot(fill = "sky blue") + theme(axis.title.x = element_blank(),
    axis.title.y = element_blank()) + ggtitle("Max Heart Rate") +
    theme(plot.title = element_text(hjust = 0.5))
taula_heart_oldpeak.bp <- ggplot(data = data.frame(taula_heart$OldPeak),
    aes(x = factor(0), y = taula_heart$OldPeak)) + theme_minimal() +
    geom_boxplot(fill = "orange") + theme(axis.title.x = element_blank(),
    axis.title.y = element_blank()) + ggtitle("Old Peak") + theme(plot.title = element_text(hjust = 0.5)
grid.arrange(taula_heart_age.bp, taula_heart_trtbps.bp, taula_heart_chol.bp,
    taula_heart_thalachh.bp, taula_heart_oldpeak.bp, nrow = 3,
    ncol = 3)
```



4. Anàlisi de les dades

Amb la funció summary fem un primer anàlisi de les dades:

```
summary(taula_heart)
```

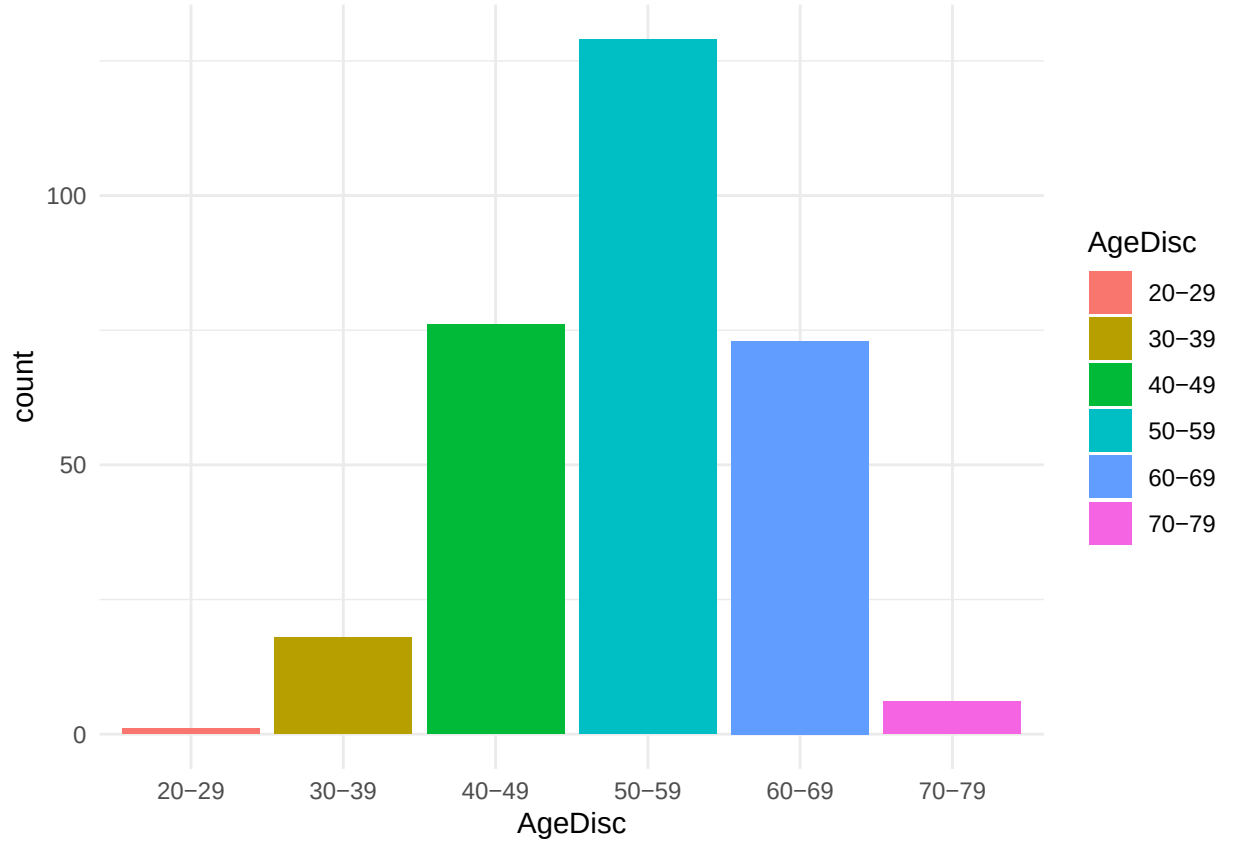
```
##       Age        Sex     ChestPain   BloodPres      Cholesterol    BloodSugar
##  Min.   :29.00   0: 96   0:143     Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50     1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87     Median :130.0   Median :240.0
##  Mean   :54.37           3: 23     Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00                     3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                     Max.   :200.0   Max.   :564.0
##     RestECG           MaxHR       ExAngina    OldPeak         Slope
##  Min.   :0.0000   Min.   : 71.0   0:204    Min.   :0.00   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:133.5   1: 99    1st Qu.:0.00   1st Qu.:1.000
##  Median :1.0000   Median :153.0            Median :0.80   Median :1.000
##  Mean   :0.5281   Mean   :149.6            Mean   :1.04   Mean   :1.399
##  3rd Qu.:1.0000   3rd Qu.:166.0            3rd Qu.:1.60   3rd Qu.:2.000
##  Max.   :2.0000   Max.   :202.0            Max.   :6.20   Max.   :2.000
##     Vessels           Thall       HeartAttack  AgeDisc
##  Min.   :0.0000   Min.   :0.000   0:138       20-29:  1
##  1st Qu.:0.0000   1st Qu.:2.000   1:165       30-39: 18
##  Median :0.0000   Median :2.000               40-49: 76
##  Mean   :0.7294   Mean   :2.314               50-59:129
##  3rd Qu.:1.0000   3rd Qu.:3.000               60-69: 73
##  Max.   :4.0000   Max.   :3.000               70-79:  6
```

- El pacient més jove té 29 anys i el més gran 77. És a dir, només s'ha realitzat l'estudi en pacients adult i la mitjana és de 54 anys.
- S'han registrat un total de 207 pacients homes i 96 pacients dones. Això ens fa pensar que el conjunt de dades no està gent equilibrat i és probable que els resultats finals de predicció de probabilitat de patir malalties cardiovasculars seran més precisos en homes que en dones al tenir-ne més mostres.
- El 47% dels pacients no han tingut cap tipus de dolor al pit.
- Un total de 165 pacients han patit un atac de cor i 138 no, el que representa un 55% vs. 45% aproximadament.
- En l'atribut del colesterol observem un valor màxim de 564.0, segurament es tracti d'un valor extrem dels que s'han vist anteriorment.

A banda de comentar estadísticament les dades veient-ne un resum, també podem observar visualitzacions ràpides per completar aquest primer anàlisi del joc de dades.

Podem observar, per exemple, un histograma amb les diferents franges d'edat i veure que on trobem més pacients és entre els 50 i 59 anys.

```
ggplot(data.frame(taula_heart), aes(x = AgeDisc, fill = AgeDisc)) +
    geom_bar() + theme_minimal()
```

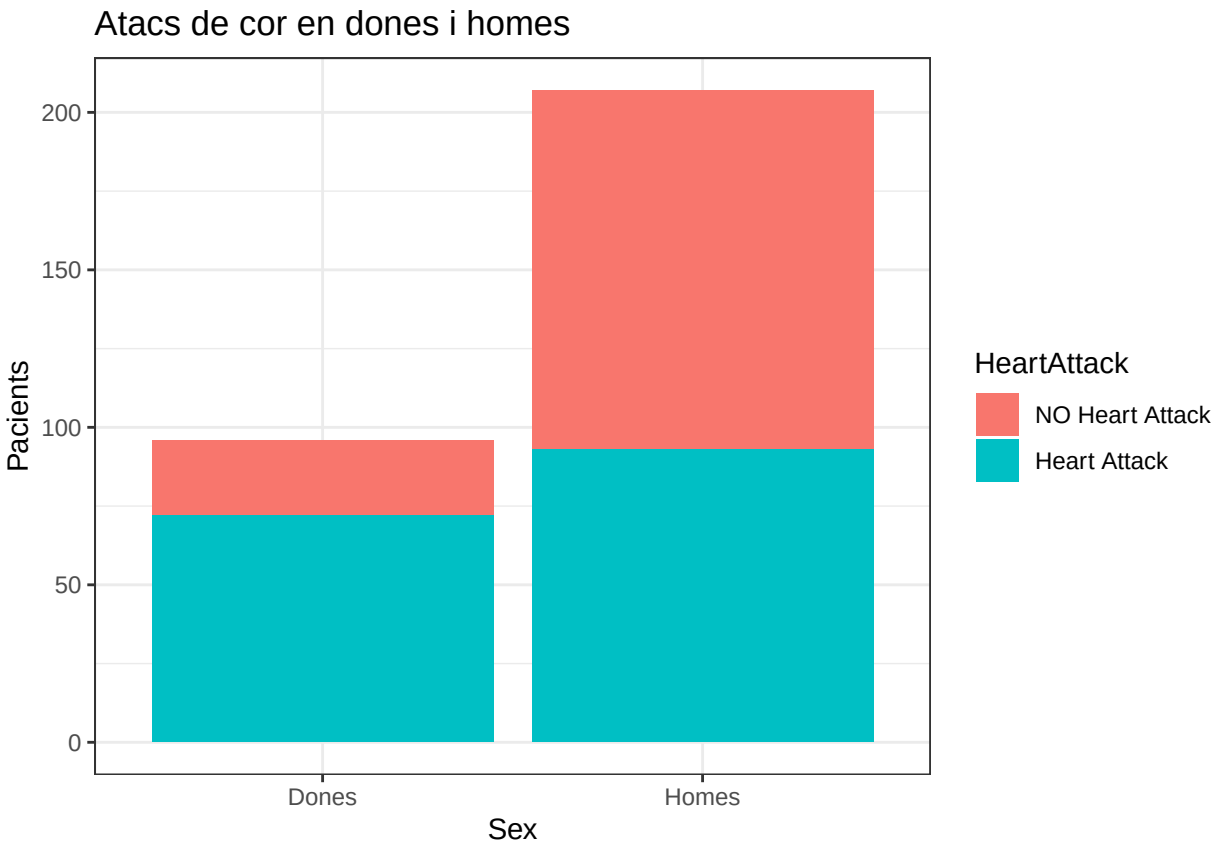


També podem comparar quin és el sexe que més atacs de cor pateix on: * Sex -> 0: Dona, 1: Home * HeartAttack -> 0: No, 1: Atac de cor

```
table(taula_heart$Sex, taula_heart$HeartAttack)
```

```
##
##       0   1
##   0  24  72
##   1 114  93
```

```
ggplot(taula_heart, aes(x = Sex, fill = HeartAttack)) + geom_bar() +
    theme_bw() + labs(y = "Pacients", title = "Atacs de cor en dones i homes") +
    scale_x_discrete(labels = c("Dones", "Homes")) + scale_fill_discrete(labels = c("NO Heart Attack",
    "Heart Attack"))
```



Amb el test de Shapiro Wilk podem revisar si les variables numèriques estan normalitzades.

```
shapiro.test(taula_heart$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  taula_heart$Age
## W = 0.98637, p-value = 0.005798
```

```
shapiro.test(taula_heart$BloodPres)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  taula_heart$BloodPres
## W = 0.96592, p-value = 1.458e-06
```

```
shapiro.test(taula_heart$Cholesterol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  taula_heart$Cholesterol
## W = 0.94688, p-value = 5.365e-09
```

```
shapiro.test(taula_heart$MaxHR)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  taula_heart$MaxHR
## W = 0.97632, p-value = 6.621e-05
```

```
shapiro.test(taula_heart$OldPeak)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  taula_heart$OldPeak
## W = 0.84418, p-value < 2.2e-16
```

Amb el test de Breush-Pagan podem saber si hi ha homogeneïtat de variança entre dos grups.

```
bptest(lm(Age ~ BloodPres, data = taula_heart))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(Age ~ BloodPres, data = taula_heart)
## BP = 4.7318, df = 1, p-value = 0.02961
```

- S'observa que el p-valor és inferior a 5%, per tant es rebutja la H0 i no hi ha homocedasticitat.

```
bptest(lm(Age ~ Cholesterol, data = taula_heart))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(Age ~ Cholesterol, data = taula_heart)
## BP = 6.3962, df = 1, p-value = 0.01144
```

- S'observa que el p-valor és inferior a 5%, per tant es rebutja la H0 i no hi ha homocedasticitat.

```
bptest(lm(Age ~ MaxHR, data = taula_heart))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(Age ~ MaxHR, data = taula_heart)
## BP = 0.58136, df = 1, p-value = 0.4458
```

- S'observa que el p-valor és superior a 5%, per tant s'accepta la H0 i hi ha homocedasticitat.

```
bptest(lm(BloodPres ~ Cholesterol, data = taula_heart))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(BloodPres ~ Cholesterol, data = taula_heart)
## BP = 2.1825, df = 1, p-value = 0.1396
```

- S'observa que el p-valor és superior a 5%, per tant s'accepta la H0 i hi ha homocedasticitat.

```
bptest(lm(BloodPres ~ MaxHR, data = taula_heart))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(BloodPres ~ MaxHR, data = taula_heart)
```

```
## BP = 0.51892, df = 1, p-value = 0.4713
```

- S'observa que el p-valor és superior a 5%, per tant s'accepta la H0 i hi ha homocedasticitat.

```
bptest(lm(Cholesterol ~ MaxHR, data = taula_heart))
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm(Cholesterol ~ MaxHR, data = taula_heart)
## BP = 0.065299, df = 1, p-value = 0.7983
```

- S'observa que el p-valor és superior a 5%, per tant s'accepta la H0 i hi ha homocedasticitat.

CONTRAST D'HIPÒTESIS:

A continuació s'aplica el model de regressió lineal que té com a objectiu aproximar la relació de dependència lineal entre una variable dependent i una (o una sèrie) de variables independents.

La regressió lineal s'aplica amb la funció *lm()*.

Veiem que la parella de variables *Age* i *BloodPres* obtenen un valor molt petit de R-squared, el que vol dir que aquestes dues variables estan molt poc correlacionades.

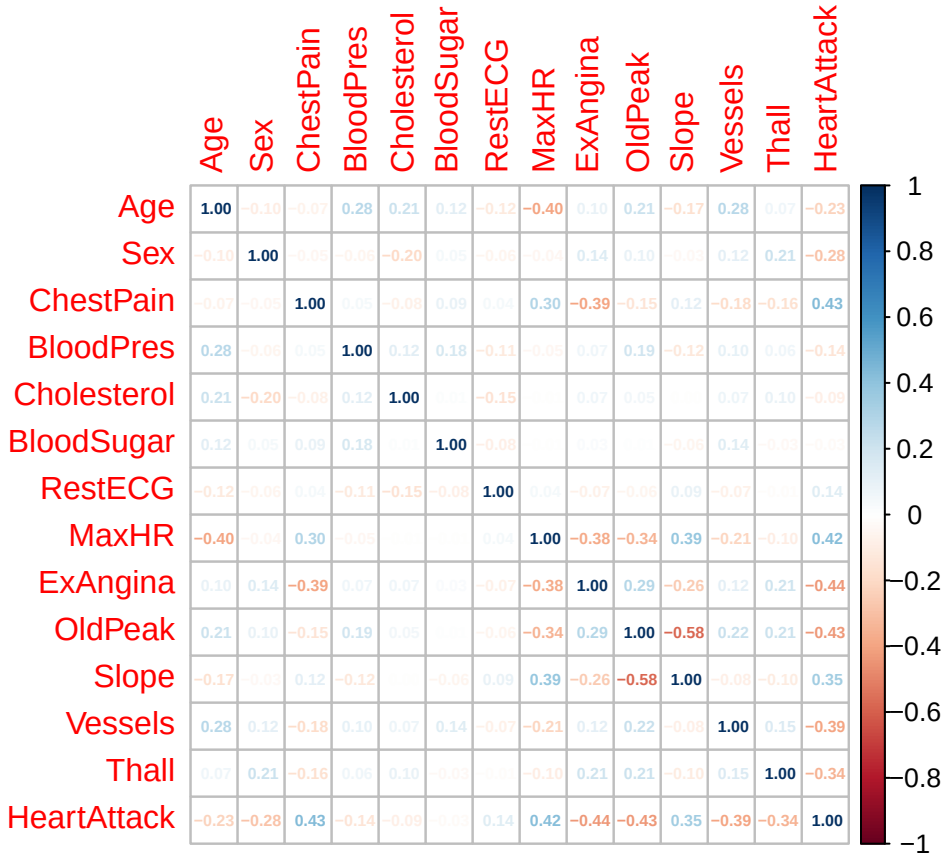
```
rl1 <- lm(Age ~ BloodPres, data = taula_heart)
summary(rl1)
```

```
##
## Call:
## lm(formula = Age ~ BloodPres, data = taula_heart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1314  -6.1441   0.5792   6.3559  23.5919
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.32545    3.80557   9.283  < 2e-16 ***
## BloodPres    0.14466    0.02866   5.048 7.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.735 on 301 degrees of freedom
## Multiple R-squared:  0.07804,    Adjusted R-squared:  0.07497
## F-statistic: 25.48 on 1 and 301 DF,  p-value: 7.762e-07
```

Apliquem a continuació una matriu de correlació sobre les diferents variables numèriques del conjunt de dades per veure quines són les que estan més relacionades entre elles, fer-ne una selecció per a posteriorment construir un algoritme de regressió que ens permeti predir la variable final *HeartAttack*.

```
taula_heart$Sex <- as.numeric(taula_heart$Sex)
taula_heart$ChestPain <- as.numeric(taula_heart$ChestPain)
taula_heart$BloodSugar <- as.numeric(taula_heart$BloodSugar)
taula_heart$ExAngina <- as.numeric(taula_heart$ExAngina)
taula_heart$HeartAttack <- as.numeric(taula_heart$HeartAttack)

corr <- cor(select_if(taula_heart, is.numeric))
corrplot(corr, method = "number", number.cex = 0.5)
```



Veiem que les variables més relacionades amb *HeartAttak* són: *MaxHR* ExAngina *OldPeak* ChestPain

I que les variables amb més correlació són *Slope* i *OldPeak*, que si analitzem la seva R-squared, obtenim un valor del 33%

```
rl2 <- lm(Slope ~ OldPeak, data = taula_heart)
summary(rl2)
```

```
##
## Call:
## lm(formula = Slope ~ OldPeak, data = taula_heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7180 -0.3597  0.2016  0.2820  1.5081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.71800    0.03889   44.18   <2e-16 ***
## OldPeak     -0.30652    0.02497  -12.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5039 on 301 degrees of freedom
## Multiple R-squared:  0.3335, Adjusted R-squared:  0.3313
## F-statistic: 150.6 on 1 and 301 DF,  p-value: < 2.2e-16
```

Així doncs, elaborem un algoritme de regressió logistica amb el conjunt de dades i la variable *HeartAttack* com

a variable dicotòmica dependent. Abans però, separem les dades en dos conjunts diferents, d'entrenament i de test.

```
split1 <- sample(c(rep(0, 0.7 * nrow(taula_heart)), rep(1, 0.3 *
    nrow(taula_heart))))
table(split1)
```

```
## split1
##   0   1
## 212  90
```

```
heart_train <- taula_heart[split1 == 0, ]
heart_test <- taula_heart[split1 == 1, ]
```

Si executem l'algoritme de regressió logística amb els conjunts de test i d'entrenament, finalment acabem obtenint una AUC (Area Under Curve) de 0.89, equivalent al rendiment del model.

```
set.seed(1234)
taula_heart$HeartAttack <- as.factor(taula_heart$HeartAttack)
glm.model <- glm(HeartAttack ~ ., data = taula_heart, family = binomial)
summary(glm.model)
```

```
##
## Call:
## glm(formula = HeartAttack ~ ., family = binomial, data = taula_heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4394  -0.3666   0.1379   0.5613   2.5991
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   20.926585 882.750272   0.024 0.981087
## Age           -0.144801   0.069579  -2.081 0.037425 *
## Sex           -1.661409   0.471779  -3.522 0.000429 ***
## ChestPain      0.865142   0.188795   4.582 4.60e-06 ***
## BloodPres     -0.017479   0.010761  -1.624 0.104319
## Cholesterol   -0.005236   0.003864  -1.355 0.175422
## BloodSugar    -0.042671   0.550472  -0.078 0.938212
## RestECG        0.425112   0.360814   1.178 0.238715
## MaxHR          0.023092   0.010679   2.162 0.030590 *
## ExAngina      -1.133371   0.432770  -2.619 0.008822 **
## OldPeak       -0.554883   0.219655  -2.526 0.011532 *
## Slope          0.499518   0.358639   1.393 0.163675
## Vessels       -0.798008   0.198892  -4.012 6.01e-05 ***
## Thall         -0.857972   0.300545  -2.855 0.004307 **
## AgeDisc30-39 -10.481653 882.744065  -0.012 0.990526
## AgeDisc40-49  -9.352593 882.744347  -0.011 0.991547
## AgeDisc50-59  -7.766222 882.745594  -0.009 0.992980
## AgeDisc60-69  -6.487106 882.747201  -0.007 0.994137
## AgeDisc70-79  -3.425322 882.751557  -0.004 0.996904
## ---
## 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: 205.81  on 284  degrees of freedom
## AIC: 243.81
##
## Number of Fisher Scoring iterations: 13
```

```
glm.model <- stats::step(glm.model, direction = "both")  # for variable selection
```

```
## Start:  AIC=243.81
## HeartAttack ~ Age + Sex + ChestPain + BloodPres + Cholesterol +
##     BloodSugar + RestECG + MaxHR + ExAngina + OldPeak + Slope +
##     Vessels + Thall + AgeDisc
##
##               Df Deviance    AIC
## - AgeDisc      5   211.44 239.44
## - BloodSugar   1   205.82 241.82
## - RestECG      1   207.21 243.21
## - Cholesterol  1   207.63 243.63
## - Slope        1   207.71 243.71
## <none>             205.81 243.81
## - BloodPres    1   208.50 244.50
## - Age          1   210.27 246.27
## - MaxHR        1   210.75 246.75
## - ExAngina     1   212.71 248.71
## - OldPeak      1   212.76 248.76
## - Thall        1   214.18 250.18
## - Sex          1   220.04 256.05
## - Vessels      1   223.33 259.33
## - ChestPain    1   230.21 266.21
##
## Step:  AIC=239.44
## HeartAttack ~ Age + Sex + ChestPain + BloodPres + Cholesterol +
##     BloodSugar + RestECG + MaxHR + ExAngina + OldPeak + Slope +
##     Vessels + Thall
##
##               Df Deviance    AIC
## - BloodSugar   1   211.44 237.44
## - Age          1   211.48 237.48
## - Cholesterol  1   212.91 238.91
## - RestECG      1   213.24 239.24
## <none>             211.44 239.44
## - Slope        1   214.13 240.13
## - BloodPres    1   215.08 241.08
## - MaxHR        1   216.64 242.64
## - ExAngina     1   217.09 243.09
## + AgeDisc      5   205.81 243.81
## - OldPeak      1   218.24 244.24
## - Thall        1   221.36 247.36
## - Sex          1   227.77 253.77
## - Vessels      1   228.84 254.84
## - ChestPain    1   236.18 262.18
##
## Step:  AIC=237.44
## HeartAttack ~ Age + Sex + ChestPain + BloodPres + Cholesterol +
##     RestECG + MaxHR + ExAngina + OldPeak + Slope + Vessels +
```

```
##     Thall
##
##               Df Deviance    AIC
## - Age          1   211.48 235.48
## - Cholesterol  1   212.91 236.91
## - RestECG      1   213.24 237.24
## <none>             211.44 237.44
## - Slope        1   214.14 238.14
## - BloodPres    1   215.10 239.10
## + BloodSugar   1   211.44 239.44
## - MaxHR        1   216.67 240.67
## - ExAngina     1   217.11 241.11
## + AgeDisc      5   205.82 241.82
## - OldPeak      1   218.33 242.33
## - Thall        1   221.80 245.80
## - Sex          1   227.90 251.90
## - Vessels      1   229.09 253.09
## - ChestPain    1   236.84 260.84
##
## Step:  AIC=235.48
## HeartAttack ~ Sex + ChestPain + BloodPres + Cholesterol + RestECG +
##     MaxHR + ExAngina + OldPeak + Slope + Vessels + Thall
##
##               Df Deviance    AIC
## - Cholesterol  1   213.08 235.08
## - RestECG      1   213.34 235.34
## <none>             211.48 235.48
## - Slope        1   214.19 236.19
## + Age          1   211.44 237.44
## + BloodSugar   1   211.48 237.48
## - BloodPres    1   215.57 237.57
## - ExAngina     1   217.12 239.12
## - MaxHR        1   218.29 240.29
## - OldPeak      1   218.34 240.34
## - Thall        1   221.86 243.86
## + AgeDisc      5   210.27 244.27
## - Sex          1   227.96 249.96
## - Vessels      1   230.16 252.16
## - ChestPain    1   236.84 258.84
##
## Step:  AIC=235.08
## HeartAttack ~ Sex + ChestPain + BloodPres + RestECG + MaxHR +
##     ExAngina + OldPeak + Slope + Vessels + Thall
##
##               Df Deviance    AIC
## <none>             213.08 235.08
## + Cholesterol  1   211.48 235.48
## - Slope        1   215.68 235.68
## - RestECG      1   215.77 235.77
## + Age          1   212.91 236.91
## + BloodSugar   1   213.08 237.08
## - BloodPres    1   217.38 237.38
## - ExAngina     1   218.81 238.81
## - MaxHR        1   219.37 239.37
```

```
## - OldPeak      1   220.55 240.55
## + AgeDisc      5   211.90 243.90
## - Thall        1   224.46 244.46
## - Sex          1   227.96 247.96
## - Vessels      1   231.65 251.65
## - ChestPain    1   239.20 259.20
```

```
summary(glm.model)
```

```
##
## Call:
## glm(formula = HeartAttack ~ Sex + ChestPain + BloodPres + RestECG +
##     MaxHR + ExAngina + OldPeak + Slope + Vessels + Thall, family = binomial,
##     data = taula_heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6203  -0.4033   0.1629   0.5876   2.5348
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.821678   2.122655   1.800 0.071794 .
## Sex         -1.563590   0.432720  -3.613 0.000302 ***
## ChestPain    0.870475   0.182131   4.779 1.76e-06 ***
## BloodPres   -0.020406   0.009988  -2.043 0.041046 *
## RestECG      0.553402   0.339071   1.632 0.102656
## MaxHR        0.022809   0.009395   2.428 0.015185 *
## ExAngina    -0.970656   0.403365  -2.406 0.016111 *
## OldPeak     -0.558608   0.211713  -2.639 0.008327 **
## Slope        0.564616   0.346525   1.629 0.103236
## Vessels     -0.763038   0.185060  -4.123 3.74e-05 ***
## Thall       -0.940441   0.284072  -3.311 0.000931 ***
## ---
## 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: 213.08  on 292  degrees of freedom
## AIC: 235.08
##
## Number of Fisher Scoring iterations: 6
```

```
glm.pred <- predict(glm.model, newdata = heart_test)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
rocobj <- roc(heart_test$HeartAttack, glm.pred, auc = TRUE)
```

```
## Setting levels: control = 1, case = 2
```

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

```
rocobj$auc
```

```
## Area under the curve: 0.9363
```

A continuació s'aplica un random forest, algoritme de classificació, amb l'objectiu d'analitzar la bondat del model amb les dades ja dividides entre test i entrenament. Amb la funció predict() es prediu el resultat de les dades del subconjunt de test i es representen les diferents mesures de bondat del model, mitjançant la funció confusionMatrix(), especificant com a positius els casos d'atac de cor.

```
taula_heart$HeartAttack <- as.factor(taula_heart$HeartAttack)
levels(taula_heart$HeartAttack)[levels(taula_heart$HeartAttack) ==
    1] <- "No Heart Attack"
levels(taula_heart$HeartAttack)[levels(taula_heart$HeartAttack) ==
    2] <- "Heart Attack"
```

```
split2 <- sample(c(rep(0, 0.7 * nrow(taula_heart)), rep(1, 0.3 *
    nrow(taula_heart))))
```

```
heart_train2 <- taula_heart[split2 == 0, ]
heart_test2 <- taula_heart[split2 == 1, ]
```

```
ha <- taula_heart[complete.cases(taula_heart), -1]
train_control <- trainControl(method = "cv", number = 4)
mod <- train(HeartAttack ~ ., data = heart_train2, method = "rf",
    trControl = train_control)
```

```
pred <- predict(mod, newdata = heart_test2)
confusionMatrix(pred, heart_test2$HeartAttack, positive = "Heart Attack")
```

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        No Heart Attack Heart Attack
##   No Heart Attack              26            6
##   Heart Attack                 15           44
##
##                Accuracy : 0.7692
##                  95% CI : (0.6691, 0.8511)
##     No Information Rate : 0.5495
##     P-Value [Acc > NIR] : 1.134e-05
##
##                   Kappa : 0.5245
##
##  Mcnemar's Test P-Value : 0.08086
##
##             Sensitivity : 0.8800
##             Specificity : 0.6341
##          Pos Pred Value : 0.7458
##          Neg Pred Value : 0.8125
##              Prevalence : 0.5495
##          Detection Rate : 0.4835
##    Detection Prevalence : 0.6484
##       Balanced Accuracy : 0.7571
```

```
##
##        'Positive' Class : Heart Attack
##
```

5. Conclusions

Generem el dataset final de sortida

```
write.csv(taula_heart, "heart_final.csv")
```

Amb tots els anàlisis fets anteriorment, es poden presentar les següents conclusions sobre el conjunt de dades:

- Les dones són més propences a patir un atac de cor
- El colesterol és una variable que va relacionada positivament a l'edat del pacient.
- Amb el simptoma dolor al pit (ChestPain), les probabilitats de patir un atac de cor augmenten gairebé en un 45%.
- Haver registrat una freqüència cardíaca màxima alta, fa que augmentin les probabilitats de patir un atac de cor.

Contribucions	**Signatura**
Investigació previa	F.V. / P.C.
Redacció de les respostes	F.V. / P.C.
Desenvolupament del codi	F.V. / P.C.
Participació al vídeo	F.V. / P.C.