PRA2. Tipologia i cicle de vida de les dades

Autor: Daniel Rodrigálvarez Morente

Maig 2023

Contents

1 Presentació del projecte i objectiu de l'anàlisi

El projecte que es desenvolupa a continuació consisteix en l'estudi de les causes que determinen la possibilitat de patir una cardiopatia. En concret, es vol determinar si els diferents indicadors estudiats tenen una incidència diferent pels homes i les dones.

2 Consideracions referents al dataset

El dataset utilitzat conté informació de diferents indicadors mèdics de persones que han patit o no una cardiopatia.

Les dades han estat publicades per Rashik Rahman sota llicència CC0: Public Domain a www.kaggle.com i es pot accedir a les mateixes a través del següent enllaç: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

No s'han realitzat modificacions prèvies al conjunt de dades original.

3 Descripció del dataset

3.1 Perquè és important i quina pregunta/problema pretén respondre?

Carreguem el conjunt de dades i fem una revisió del contingut de les diferents variables

```
path = 'heart.csv'
dades <- read.csv(path, sep = ",")
str(dades)
```

```
## '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 ...
```

Podem observar que es tracta d'un dataset amb 303 observacions i 14 variables, totes elles amb números enters excepte la variable oldpeak que conté dades decimals.

Descripció de les variables

- **age**. Edat de la persona

- **sex**. Sexe de la persona (1 = home; 0 = dona)

- **cp**. chest pain type Value (1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic)

- **trtbps**. Pressió arterial en repòs (en mm/Hg)

- **chol**. Nivell de colesterol mesurat (en mg/dl)

- **fbs**. Nivell de sucre en sang en dejú (fasting blood sugar > 120 mg/dl) (1: > 120 mg/dl; 0: =< 120 mg/dl)

- **restecg**. resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)

- **thalachh**. Freqüència cardiaca màxima assolida

- **exng**. Angina induïda per exercici físic (1 = sí; 0 = no)

- **oldpeak**. Previous peak

- **slp**. Slope (0 = unsloping 1 = flat 2 = downsloping)

- **caa**. Número de vasos sanguinis principals amb obstrucció (0-4)

- **thall**. Resultats d'una prova d'esforç amb tali (0 = null; 1 = fixed defect; 2 = normal; 3 = reversable defect)

- **output**. Variable objectiu (0 = menys possibilitats de partir una cardiopatia (< 50% diameter narrowing. less chance of heart disease); 1= més possiblitats de patir una cardiopatia (> 50% diameter narrowing. more chance of heart disease))

4 Integració i selecció de les dades d'interès a analitzar. Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.

Revisem la distribució de les diferents variables

```
summary(dades)
```

```
##       age             sex               cp            trtbps     
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##       chol            fbs             restecg          thalachh     
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##       exng           oldpeak          slp             caa        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
##      thall           output     
##  Min.   :0.000   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:0.0000  
##  Median :2.000   Median :1.0000
```

```
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

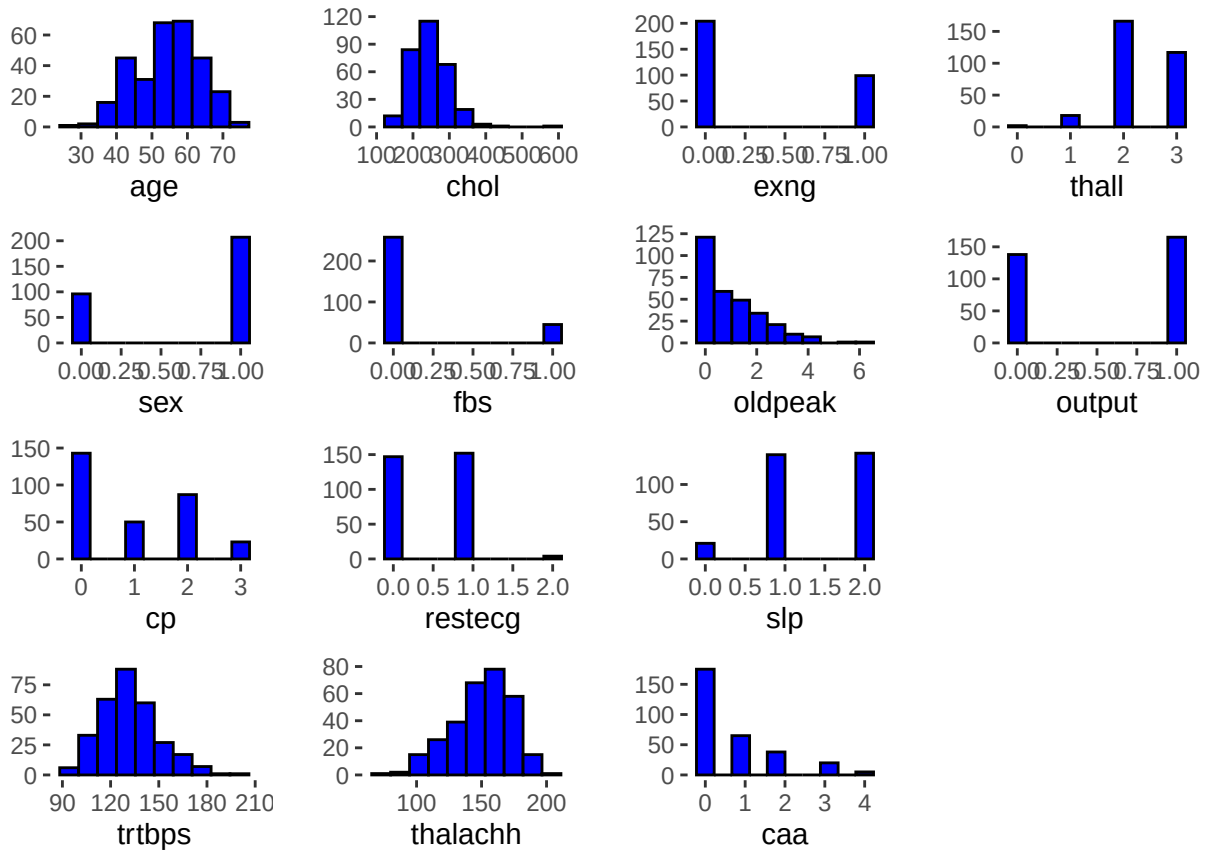
A priori, no observem que hi hagi valor perduts, però més endevant farem una comprovació adhoc.

Fem una representació de les diferents variables per tal de facilitar la revisió prèvia del dataset

```
histogrames_num <- list()
variables_num <- names(dades)
dades_num <- dades %>% select(all_of(variables_num))

for(i in 1:ncol(dades_num)){
  var <- names(dades_num)[i]
  grafic <- ggplot(dades_num, aes_string(x = var)) +
    geom_histogram(bins = 10, fill = "blue", color = "black") +
    labs(y = "") +
    theme(panel.grid = element_blank(), panel.background = element_blank())
  histogrames_num[[i]] <- grafic
}

multiplot(plotlist = histogrames_num, cols = 4)
```



Hi ha més informació d'homes que de dones i el número de registres amb output igual a 1 és lleugerament superior al valor 0.

Observant les gràfiques, veiem que hi ha quatre variables que podrien tenir una distribució similar a una

normal (age, chol, trtbps i thalachh). En tot cas, més endavant farem una comprovació adhoc per tal d'assegurar-ho.

Modificarem els valors de la variable sex per facilitar la seva interpretació

```
dades$sex[dades$sex == 1] <- "Home"
dades$sex[dades$sex == 0] <- "Dona"
```

Revisem si tenim registres amb identics valors a totes les variables per tal de valorar si tenim registres duplicats

```
dim(unique(dades))
```

```
## [1] 302  14
```

Comprovem que hi ha 302 registres diferents, per la qual cosa, donat el nivell d'especificitat de les dades, considerem que hi ha un registre repetit.

Eliminem el registre repetit i conservem la resta donat que tenim un número de registres perfectement gestionable i, per tant no és necessari plantejar agrupacions que facilitin l'ús del dataset

```
dades <- unique(dades)
```

Donat que l'estudi està dirigit a identificar diferències entre homes i dones, ens interessa comprobar quina informació tenim per cada grup

```
print('Distribució entre homes i dones en valors absoluts:')
```

```
## [1] "Distribució entre homes i dones en valors absoluts:"
```

```
print(addmargins(table(dades$sex, dades$output)))
```

```
##
##          0   1 Sum
##   Dona  24  72  96
##   Home 114  92 206
##   Sum  138 164 302
```

```
print('Pes relatiu de cada sexe dins el valor de la variable output:')
```

```
## [1] "Pes relatiu de cada sexe dins el valor de la variable output:"
```

```
print(round(prop.table(table(dades$sex, dades$output), 2), 2))
```

```
##
##           0    1
##   Dona 0.17 0.44
##   Home 0.83 0.56
```

```r
print('Pes relatiu de la variable output dins de cada sexe:')
```

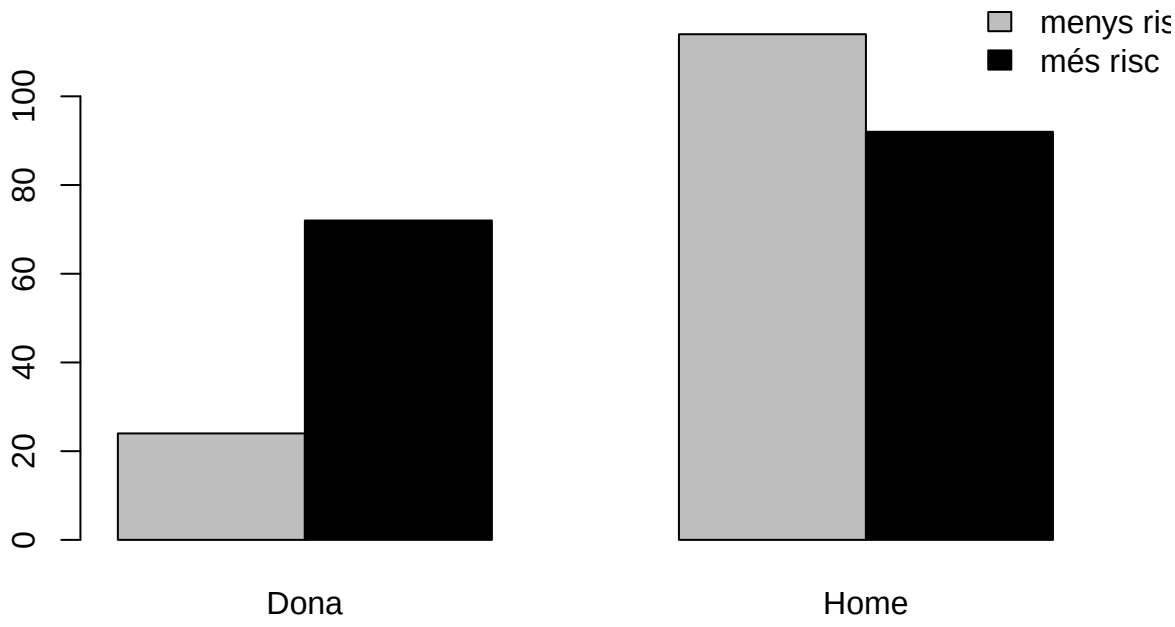
```
## [1] "Pes relatiu de la variable output dins de cada sexe:"
```

```r
print(round(prop.table(table(dades$sex, dades$output), 1), 2))
```

```
##
##            0    1
##   Dona 0.25 0.75
##   Home 0.55 0.45
```

Hi ha 207 homes i 96 dones i la distribució del camp output dins de cada grup és diferent, tenint més pes el valor 1 en dones que en homes.

```r
grafic <- barplot(table(dades$output, dades$sex),
                  beside = TRUE,
                  col = c("grey", "black"), legend = FALSE)
legend("topright", legend = c("menys risc", "més risc"),
       fill = c("grey", "black"),
       x = max(grafic),
       y = max(grafic) + 120,
       xpd = TRUE,
       bty = "n")
```



5 Neteja de les dades. Les dades contenen zeros o elements buits?

5.1 Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos

Tot i que amb el resum del dataset no apareixien valors perduts, fem una adhoc

```
colSums(is.na(dades))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

```
colSums(dades=="")
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

Es confirma que no tenim valors perduts, per tant no em de fer cap modificació al dataset.

5.2 Identifica i gestiona els valors extrems

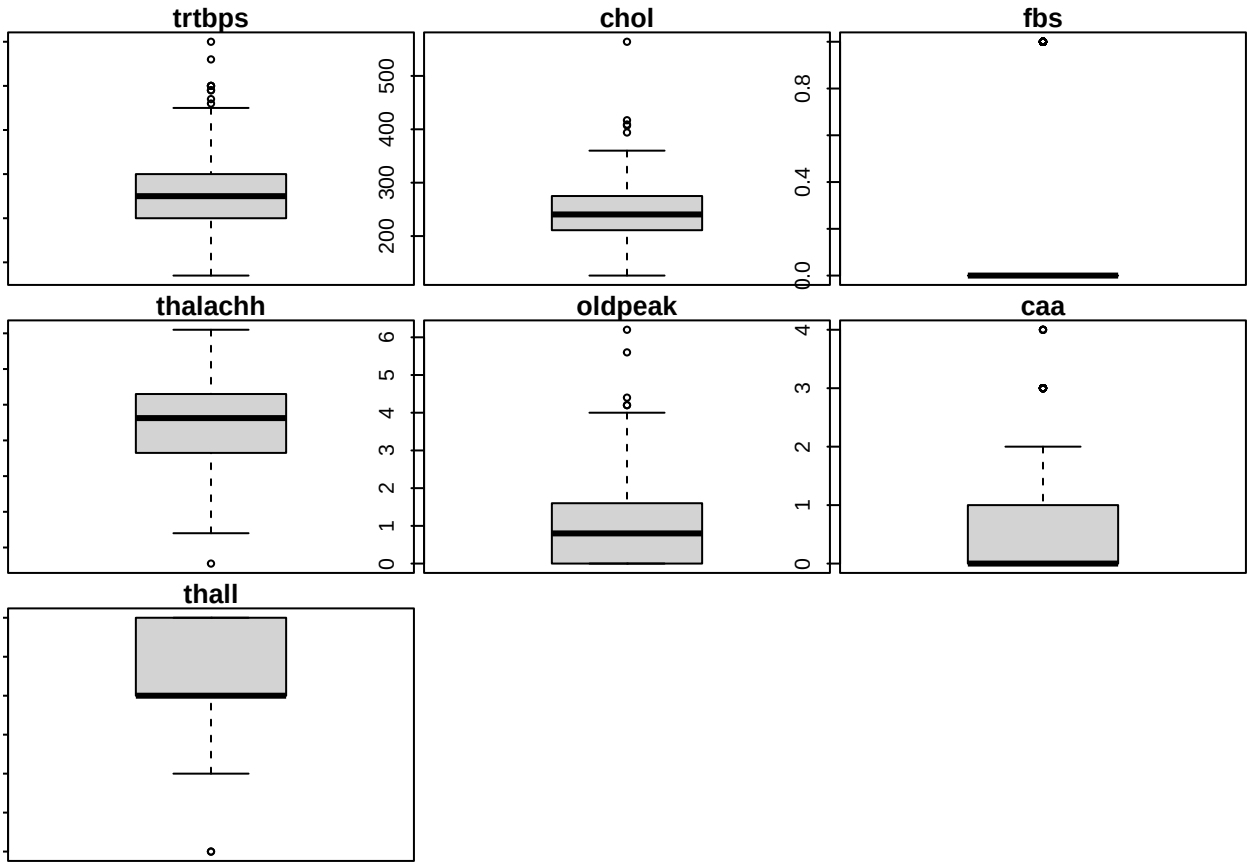
Fem una revisió de la possible existència de valors extrems de forma global

```
var_out <- c()
for (i in c(1,3:ncol(dades))){
  outl <- boxplot.stats(dades[,i])$out
  if (!length(outl)==0){var_out <- c(var_out, i)}
}
print(names(dades)[var_out])
```

```
## [1] "trtbps"   "chol"     "fbs"      "thalachh" "oldpeak"  "caa"      "thall"
```

Veiem que tenim 7 variables amb valors extrems

```
par(mfrow = c(3,3), mar = c(0, 0, 1, 0) + 0.2)
for (i in var_out) {
  boxplot(dades[, i], main = colnames(dades)[i])
}
```



Per les variables fbs, caa i thall, tot i el resultat del gràfics, podem descartar el fet que hi hagi valors extrems donat que es tracta de variables discretes i els valors observats estan dins de les categories considerades.

Apliquem el criteri de les dues desviacions estàndar per tal de valorar si mantenim el valors originals

```
trtbps_outliers <- abs(scale(dades$trtbps)) > 2
chol_outliers <- abs(scale(dades$chol)) > 2
thalachh_outliers <- abs(scale(dades$thalachh)) > 2
oldpeak_outliers <- abs(scale(dades$oldpeak)) > 2

outliers <- trtbps_outliers + chol_outliers + thalachh_outliers + oldpeak_outliers
print(paste("Número d'outliers considerant les dades de forma global:",
            count(dades[outliers > 0,])))
```

```
## [1] "Número d'outliers considerant les dades de forma global: 49"
```

Donat que l'estudi es basa en la comparació entre homes i dones, fe una valoració dels valors extrems per separat, per tal d'evitar que els valors d'un sexe amaguin informació rellevant a l'altra

```
dades_homes <- dades[dades$sex=="Home",]
dades_dones <- dades[dades$sex=="Dona",]

var_out_dones <- c()
for (i in c(1,3:ncol(dades))){
  outl_dones <- boxplot.stats(dades_dones[,i])$out
  if (!length(outl_dones)==0){var_out_dones <- c(var_out_dones, i)}
```

```
}
print(names(dades_dones)[var_out_dones])
```

```
## [1] "trtbps"   "chol"     "fbs"      "thalachh" "exng"     "oldpeak"  "caa"
## [8] "thall"
```

```
var_out_homes <- c()
for (i in c(1,3:ncol(dades))){
  outl_homes <- boxplot.stats(dades_homes[,i])$out
  if (!length(outl_homes)==0){var_out_homes <- c(var_out_homes, i)}
}
print(names(dades_homes)[var_out_homes])
```

```
## [1] "trtbps"   "fbs"      "thalachh" "oldpeak"  "caa"      "thall"
```

Així com en l'estudi conjunt trobavem valors extrems a les variables trtbps, chol, thalachh i oldpeak, quan estudiem els sexes per separat varien aquests resultats. Per les dones no hi ha variació en quant a variables, donat que la variable exng en ser discreta no la podem considerar, i pels homes no hi haurà valors extrems per la variable chol.

Revisem els registres que contenen valors extrems considerant dues desviacions estàndar i comptem quants registres es veuen afectats per outliers, considerant les dades de forma global i separant per sexes

```
print(paste("Número d'outliers considerant les dades de forma global:",
            count(dades[outliers > 0,])))
```

```
## [1] "Número d'outliers considerant les dades de forma global: 49"
```

```
print(paste("Número d'outliers en dones considerant les dades de forma global:",
            count(dades[c(outliers > 0 & dades$sex=="Dona"),])))
```

```
## [1] "Número d'outliers en dones considerant les dades de forma global: 19"
```

```
print(paste("Número d'outliers en homes considerant les dades de forma global:",
            count(dades[c(outliers > 0 & dades$sex=="Home"),])))
```

```
## [1] "Número d'outliers en homes considerant les dades de forma global: 30"
```

```
trtbps_outliers_d <- abs(scale(dades_dones$trtbps)) > 2
chol_outliers_d <- abs(scale(dades_dones$chol)) > 2
thalachh_outliers_d <- abs(scale(dades_dones$thalachh)) > 2
oldpeak_outliers_d <- abs(scale(dades_dones$oldpeak)) > 2

outliers_dones <- trtbps_outliers_d + chol_outliers_d + thalachh_outliers_d + oldpeak_outliers_d
print(paste("Número d'outliers en dones:", count(dades_dones[outliers_dones > 0,])))
```

```
## [1] "Número d'outliers en dones: 17"
```

```r
trtbps_outliers_h <- abs(scale(dades_homes$trtbps)) > 2
thalachh_outliers_h <- abs(scale(dades_homes$thalachh)) > 2
oldpeak_outliers_h <- abs(scale(dades_homes$oldpeak)) > 2

outliers_homes <- trtbps_outliers_h + thalachh_outliers_h + oldpeak_outliers_h
print(paste("Número d'outliers en homes:", count(dades_homes[outliers_homes > 0,])))
```

```
## [1] "Número d'outliers en homes: 25"
```

Podem extreure una primera conclusió sobre la importància de tractar les dades per separat donat que es reduix el número d'outliers. Tractar les dades conjuntament implicaria fer un tractament de les dades errònia i descartar registres o imputar valors de forma equivocada, a banda de que ens facilita una primera informació al respecte dels diferents valors observats en funció de si es tracta de dones o d'homes.

Per tal de decidir si realment els valors trobats són erronis cal tenir un coneixement ampli del tipus de dades i de si els valors que estem identificant són relament erronis. Per altra banda, amb una mostra de 302 registres si 41 tenen dades errònies, hauríem de considerar que hi ha hagut massa errors en la recolecció de les dades i la mostra no és gaire útil. Per tant, considerarem que valors detectats són correctes i continuarem l'anàlisi sense imputar nous valors, tenint present que aquest és un exercici teòric i que en un cas real hauríem de consultar amb els experts per tal de validar quin és el tractament correcte.

6 Anàlisi de les dades

6.1 Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

Com ja s'ha comentat, l'objectiu de l'estudi és la comparació de les dades existents entre homes i dones per tal de valorar la diferent incidència que puguin tenir els resultats de les proves en el fet de patir una cardiopatia.

Ja hem vist que el tractament diferenciat ens porta a detectar valors extrems diferents en el cas d'homes i de dones.

Tot i que es tracta d'un tècnica utilitzada per reduir la dimensionalitat de les dades, farem una anàlisi PCA per tal de valorar si les variables tenen la mateixa importància pels dos grups i d'aquesta manera refermar la idea de fer un tractament diferenciat

```r
dades_acp_dones <- prcomp(dades_dones[,c(1,3:ncol(dades_dones))], center = TRUE, scale = TRUE)
print("ACP dones:")
```

```
## [1] "ACP dones:"
```

```r
summary(dades_acp_dones)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.9353 1.2488 1.1511 1.05246 1.00397 0.95263 0.91461
## Proportion of Variance 0.2881 0.1200 0.1019 0.08521 0.07753 0.06981 0.06435
## Cumulative Proportion  0.2881 0.4081 0.5100 0.59522 0.67276 0.74257 0.80691
##                            PC8     PC9    PC10    PC11    PC12   PC13
```

```
## Standard deviation     0.87141 0.69795 0.63454 0.55813 0.54281 0.5048
## Proportion of Variance 0.05841 0.03747 0.03097 0.02396 0.02266 0.0196
## Cumulative Proportion  0.86532 0.90280 0.93377 0.95773 0.98040 1.0000
```

```
dades_acp_homes <- prcomp(dades_homes[,c(1,3:ncol(dades_homes))], center = TRUE, scale = TRUE)
print("ACP homes:")
```

```
## [1] "ACP homes:"
```

```
summary(dades_acp_homes)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7966 1.2412 1.12198 1.0268 0.99126 0.94694 0.91272
## Proportion of Variance 0.2483 0.1185 0.09683 0.0811 0.07558 0.06898 0.06408
## Cumulative Proportion  0.2483 0.3668 0.46364 0.5447 0.62032 0.68929 0.75338
##                           PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.87956 0.80828 0.76272 0.67198 0.64432 0.57508
## Proportion of Variance 0.05951 0.05026 0.04475 0.03474 0.03193 0.02544
## Cumulative Proportion  0.81289 0.86314 0.90789 0.94263 0.97456 1.00000
```

S'aprecia una diferència entre el resultat pels homes i per les dones, sent força significativa la diferència de la primera component principal, amb un valor de quasi un 4% més per les dones.

En quant a la interpretació del resultat, per les dones les dues primeres components principals expliquen el 40,81% de la variància, mentre que pels homes acumulen 36,68%, però la participació de les diferents components principals està força repartida, necessitant fins 9 per les dones i 10 pels homes per arribar al 90%.

Com a criteri de selecció considerarem les components amb una variància superior a 1

```
var_dades_acp_dones <- dades_acp_dones$sdev ^ 2
print(var_dades_acp_dones)
```

```
##  [1] 3.7455634 1.5595211 1.3251408 1.1076699 1.0079545 0.9075065 0.8365109
##  [8] 0.7593494 0.4871390 0.4026351 0.3115098 0.2946376 0.2548619
```

```
var_dades_acp_homes <- dades_acp_homes$sdev ^ 2
print(var_dades_acp_homes)
```

```
##  [1] 3.2278984 1.5405512 1.2588347 1.0542459 0.9826044 0.8966912 0.8330578
##  [8] 0.7736301 0.6533156 0.5817416 0.4515623 0.4151547 0.3307122
```

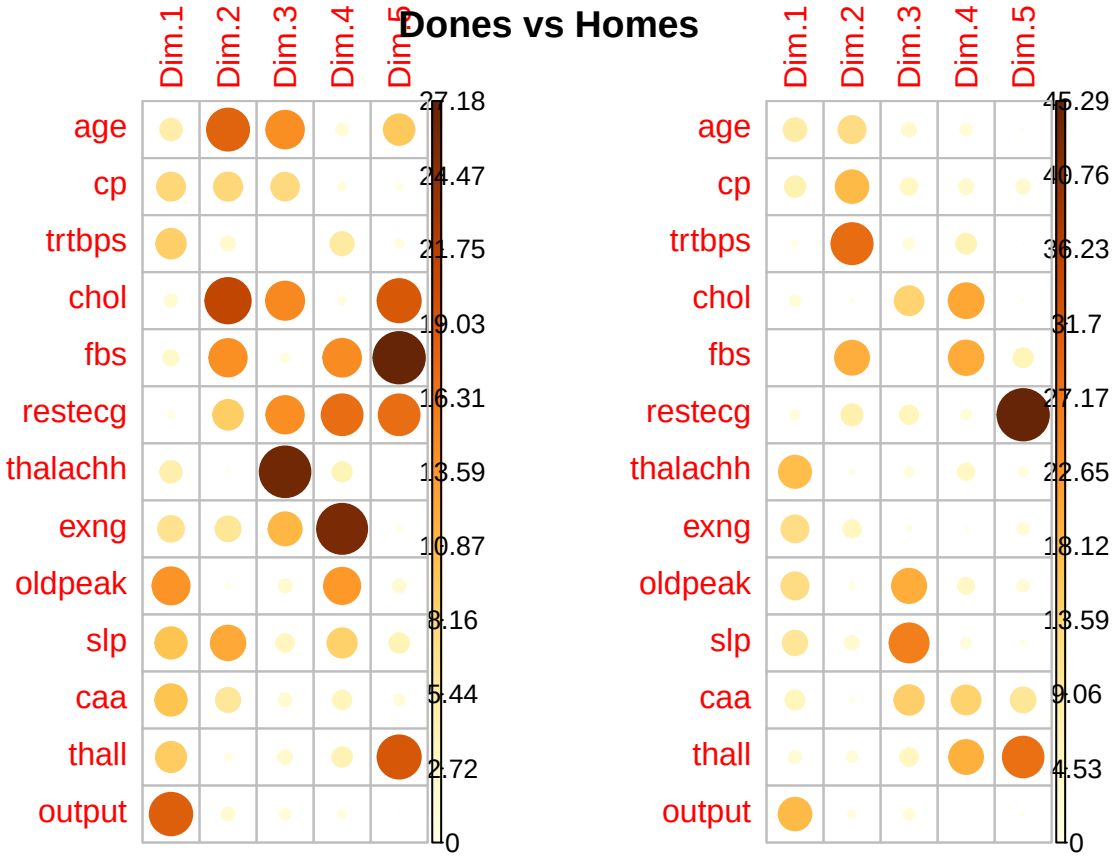
Per tant, considerarem les 5 primeres per les dones i les 4 primeres pels homes. De totes maneres, revisarem la importància de cada variable a les 5 primeres components principals diferenciant per sexes

Ara mirem com intervenen les variables en cada una de les 5 primeres components principals (tot i que per les dones en tenim prou amb 4, utilitzarem les 5 primeres per comparar els dos grups)

```r
var_dones <- get_pca_var(dades_acp_dones)
var_homes <- get_pca_var(dades_acp_homes)

par(mfrow = c(1,2))

corrplot(var_dones$contrib[,1:5], is.corr=FALSE)
corrplot(var_homes$contrib[,1:5], is.corr=FALSE)
title(main="Dones vs Homes", outer = TRUE, line = -1)
```



De forma visual es pot apreciar que hi ha diferències en quant a la importància de cada variable en la contribució a les 5 primeres components principals, per tant podem considerar que a l'hora de seleccionar les variables a estudiar serà important diferenciar entre els dos sexes.

6.2 Comprovació de la normalitat i homogeneïtat de la variància.

Tenim cinc variables contínues: age, trtbps, chol, thalachh i oldpeak a les quals aplicarem el test de Shapiro per comprovar la seva **normalitat** a les dades en conjunt

```r
shapiro.test(dades$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$age
## W = 0.98664, p-value = 0.006745
```

```r
shapiro.test(dades$trtbps)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$trtbps
## W = 0.96573, p-value = 1.419e-06
```

```r
shapiro.test(dades$chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$chol
## W = 0.94658, p-value = 5.196e-09
```

```r
shapiro.test(dades$thalachh)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$thalachh
## W = 0.97679, p-value = 8.268e-05
```

```r
shapiro.test(dades$oldpeak)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dades$oldpeak
## W = 0.84522, p-value < 2.2e-16
```

En els cinc casos, observant el valor de p, podem dir que, segons el test Shapiro, es rebutja la hipòtesi nul · la i, per tant, no es distribueixen com una normal.

Apliquem el test de Kolmogorov_Smirnov per tal de valorar si els resultats els mateixos

```r
ks.test(dades$age, pnorm, mean(dades$age), sd(dades$age))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  dades$age
## D = 0.075788, p-value = 0.06228
## alternative hypothesis: two-sided
```

```r
ks.test(dades$trtbps, pnorm, mean(dades$trtbps), sd(dades$trtbps))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  dades$trtbps
## D = 0.10258, p-value = 0.003475
## alternative hypothesis: two-sided
```

```
ks.test(dades$chol, pnorm, mean(dades$chol), sd(dades$chol))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  dades$chol
## D = 0.055822, p-value = 0.3035
## alternative hypothesis: two-sided
```

```
ks.test(dades$thalachh, pnorm, mean(dades$thalachh), sd(dades$thalachh))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  dades$thalachh
## D = 0.070819, p-value = 0.09669
## alternative hypothesis: two-sided
```

```
ks.test(dades$oldpeak, pnorm, mean(dades$oldpeak), sd(dades$oldpeak))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  dades$oldpeak
## D = 0.18458, p-value = 2.313e-09
## alternative hypothesis: two-sided
```

Obtenim resultats contradictoris per age, chol i thalachh, per tant, sent conservadors, considerarem que cap de les cinc variables es distribueix segons una normal

Comprovem l'**homoscedasticitat** per les quatre variables

Apliquem el test de fligner sobre les cinc variables, al considerar que no segueixen una distribució normal

```
fligner.test(trtbps ~ sex, data = dades)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  trtbps by sex
## Fligner-Killeen:med chi-squared = 0.93812, df = 1, p-value = 0.3328
```

```
fligner.test(oldpeak ~ sex, data = dades)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  oldpeak by sex
## Fligner-Killeen:med chi-squared = 8.372, df = 1, p-value = 0.00381
```

```
fligner.test(age ~ sex, data = dades)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  age by sex
## Fligner-Killeen:med chi-squared = 0.68171, df = 1, p-value = 0.409
```

```
fligner.test(chol ~ sex, data = dades)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chol by sex
## Fligner-Killeen:med chi-squared = 9.2927, df = 1, p-value = 0.002301
```

```
fligner.test(thalachh ~ sex, data = dades)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  thalachh by sex
## Fligner-Killeen:med chi-squared = 5.3763, df = 1, p-value = 0.02041
```

Segons els resultats obtingut podem concloure que oldpeak, chol i thalachh tenen variàncies estadísticament diferents per cada sexe mentre que age i trtbps tenen variànces estadísticamente iguals per cada sexe.

6.3 Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

En base a l'estudi de components principals realitzats prèviament, considerarem les següents variables: oldpeak, age i chol, per la incidència que tenen en el grup de dones, i thalachh, trtbps i fbs, per la incidència que tenen en el grup d'homes. A més considerarem la variable output.

Donat que hem considerat que les variables contínues no segueixen una distribució normal, utilitzarem test no paramètrics per aquests variables.

Considerem que els grups dades separats per sexe són independents per la qual cosa aplicarem el test Mann-Whitney per comparar les distribucions

```
wilcox.test(oldpeak ~ sex, data = dades)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  oldpeak by sex
## W = 8638, p-value = 0.07187
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(age ~ sex, data = dades)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by sex
## W = 11064, p-value = 0.09585
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(chol ~ sex, data = dades)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  chol by sex
## W = 11710, p-value = 0.009943
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(thalachh ~ sex, data = dades)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  thalachh by sex
## W = 10420, p-value = 0.4519
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(trtbps ~ sex, data = dades)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trtbps by sex
## W = 10552, p-value = 0.3464
## alternative hypothesis: true location shift is not equal to 0
```

Donat que obtenim una p-valor inferior al 0,05 per chol podem concloure que aquesta variable té diferències estadísticament significatives per cada sexe, mentre que a les altres quatre variables contínues no.

Apliquem el test chi-quadrat a la variable discreta fbs

```
chisq.test(table(dades[, c("sex", "fbs")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(dades[, c("sex", "fbs")])
## X-squared = 0.39221, df = 1, p-value = 0.5311
```

Podem concloure que no hi ha diferències significatives en la variable fbs per sexes.

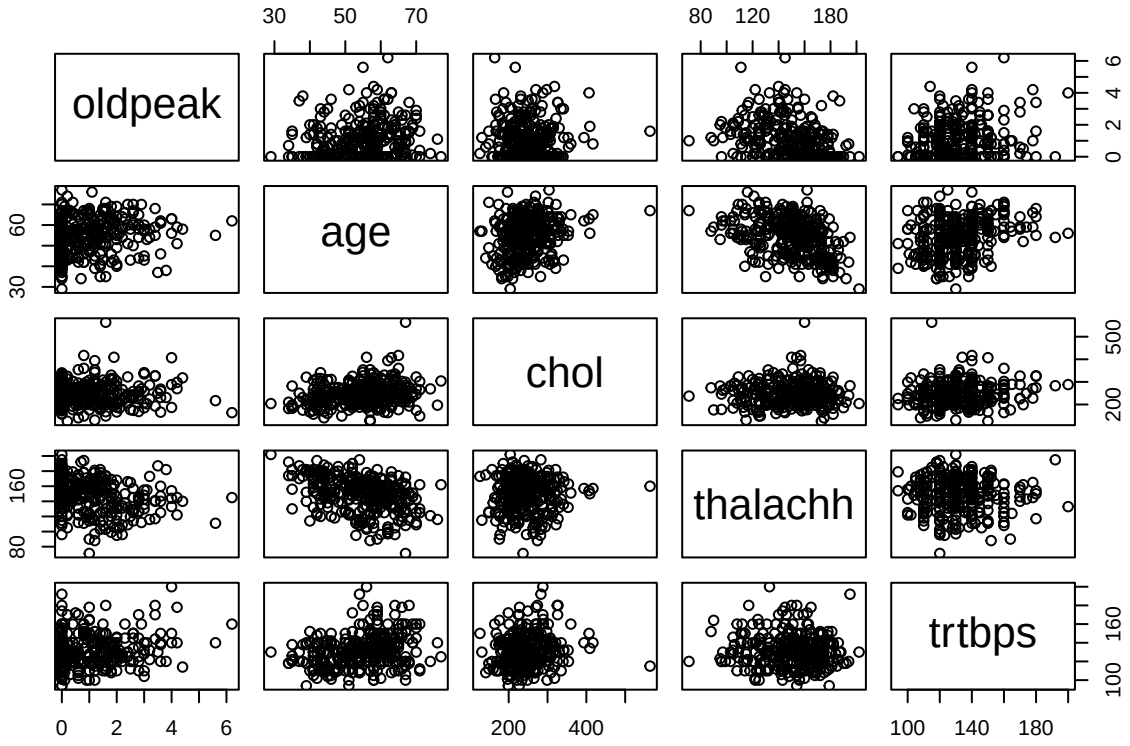
```
chisq.test(table(dades[, c("sex", "output")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(dades[, c("sex", "output")])
## X-squared = 23.084, df = 1, p-value = 1.551e-06
```

En aquest cas sí s'aprecien diferències significatives, la qual cosa és especialment rellevant pel nostre estudi, sempre i quan considerem que la mostra es suficientment representativa de la població.

Observem les relacions entre les diferents variables contínues per tal valorar si hi pot haver correlació entre elles que ens permeti generar un model de regressió lineal.

```
plot(dades[, c("oldpeak", "age", "chol", "thalachh", "trtbps")])
```



A priori, sembla que hi podria haver una certa relació entre age i chol i també entre age i thalachh

```r
age_chol <- lm(age ~ chol, dades)
summary(age_chol)
```

```
##
## Call:
## lm(formula = age ~ chol, data = dades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.8809  -6.5405   0.4127   6.3624  23.3727
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.490546   2.486992  18.291  < 2e-16 ***
## chol         0.036227   0.009875   3.669 0.000288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.866 on 300 degrees of freedom
## Multiple R-squared:  0.04294,    Adjusted R-squared:  0.03975
## F-statistic: 13.46 on 1 and 300 DF,  p-value: 0.0002884
```

```r
age_thalachh <- lm(age ~ thalachh, dades)
summary(age_thalachh)
```

```
##
## Call:
## lm(formula = age ~ thalachh, data = dades)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.4760  -6.6945   0.5537   6.3865  24.5203
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 77.77379    3.17004  24.534  < 2e-16 ***
## thalachh    -0.15614    0.02095  -7.452 9.86e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.325 on 300 degrees of freedom
## Multiple R-squared:  0.1562, Adjusted R-squared:  0.1534
## F-statistic: 55.54 on 1 and 300 DF,  p-value: 9.858e-13
```

Com podem observar en el resultats, la correlació entre els dos parells de variables és feble.

Donat que hem considerat sis variables com a més significatives per l'estudi, calcularem la correlació existent entre les cinc d'aquestes que són contínues i output per valorar el nivell de correlació existent

```r
cor(dades[, c("oldpeak", "age", "chol", "thalachh", "trtbps", "output")], method = "spearman")
```

```
##              oldpeak        age        chol    thalachh      trtbps     output
```

```
## oldpeak   1.00000000  0.2636254  0.03956479 -0.43049461  0.15680732 -0.4196306
## age       0.26362540  1.0000000  0.18890292 -0.39345342  0.28970501 -0.2348453
## chol      0.03956479  0.1889029  1.00000000 -0.04036747  0.13021023 -0.1170065
## thalachh -0.43049461 -0.3934534 -0.04036747  1.00000000 -0.04269948  0.4263680
## trtbps    0.15680732  0.2897050  0.13021023 -0.04269948  1.00000000 -0.1234777
## output   -0.41963058 -0.2348453 -0.11700649  0.42636798 -0.12347766  1.0000000
```

```r
cor.test(dades$oldpeak, dades$output, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  dades$oldpeak and dades$output
## S = 6516887, p-value = 2.604e-14
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4196306
```

7 Conclusions

A partir dels resultats obtinguts, quines són les conclusions?

Els resultats permeten respondre al problema?

8 Taula de contribucions

Contribucions	Signatura
Investigació prèvia	Daniel Rodrigálvarez Morente
Redacció de les respostes	Daniel Rodrigálvarez Morente
Desenvolupament	Daniel Rodrigálvarez Morente
Participació al vídeo	Daniel Rodrigálvarez Morente