PRA 2 - Com realitzar la neteja i l'anàlisi de dades?

Judith Cid, Jordi Tormo

2023-12-27

Contents

1. Descripció del dataset. **1**

2. Integració i selecció. **2**

3. Neteja de les dades. **2**

4. Anàlisi de les dades. **5**

5. Resolució del problema. **20**

6. Vídeo. **20**

7. Taula de contribucions. **20**

1. Descripció del dataset.

El dataset "Heart Attack Analysis & Prediction" és un conjunt de dades que conté informació pacients identificant-ne si ténen més o menys probabilitat a patir un atac de cor. Aquest dataset és important perquè pot ser utilitzat per entrenar models de predicció per preveure qui és més propens a patir un atac de cor en el futur. El dataset pot ajudar a respondre a la pregunta de quins són els factors de risc més importants per a la salut cardiovascular.

El conjunt de dades s'ha obtingut de l'enllaç proporcionat a l'enunciat (https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset). Dels dos documents disponibles a l'enllaç, es treballarà amb el document `heart.csv`.

- `age`: Edat del pacient
- `sex`: Sexe del pacient
- `cp` : Tipus de dolor toràcic.
  - `1`: Angina de pit típica
  - `2`: Angina de pit atípica
  - `3`: Dolor no anginal
  - `4`: Asimptomàtic
- `trtbps`: Pressió arterial en repòs (in mm Hg)
- `chol`: Colesterol en mg/dl obtingut mitjançant sensor d'IMC (índex de massa corporal)
- `fbs`: Sucre en sang en dejú > 120 mg/dl
  - `1`: Vertader
  - `0`: Fals
- `restecg`: Resultats electrocardiogràfics en repòs
  - `0`: Normal
  - `1`: Té anomalia de l'ona ST-T (inversions d'ona T i/o elevació o depressió ST de $> 0{,}05$ mV)

  - 2: Mostra una hipertròfia ventricular esquerra probable o definitiva segons el criteri d'Estes
- `thalachh`: Freqüència cardíaca màxima aconseguida
- `exng`: Exercici angina de pit induïda
  - 1: Sí
  - 0: No
- `oldpeak`: Depressió de ST induïda per l'exercici en relació amb el descans
- `slp`: El pendent del segment ST d'exercici màxim
  - 0: Sense inclinar
  - 1: Pla
  - 2: Baixada
- `caa`: Nombre de vasos principals del cor (0-4)
- `thall`: Talassèmia
  - 0: Nul
  - 1: Defecte solucionat
  - 2: Normal
  - 3: Defecte reversible
- `output`: Diagnòstic de malalties cardíaques (estat de malaltia angiogràfica)
  - 0: Estrenyiment $< 50\%$ del diàmetre. Menys possibilitats de patir malalties del cor.
  - 1: Estrenyiment $> 50\%$ del diàmetre. Més possibilitats de patir malalties del cor.

2. Integració i selecció.

S'utilitza el document `heart.csv` que consta de 303 registres i 14 atributs

```
# Es carrega el fitxer de dades en un objecte amb identificador denominat data
data_raw <- read.csv('../Dataset/heart.csv',header=T,sep=",")
str(data_raw)
```

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

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits?

En primer lloc, es comprova que tots els atributs tenen sentit segons la seva descripció i valors possibles.

```
# s'examinen els valors resum de cada tipus de variable
data <- data_raw
```

```
print(paste("CP: A part dels valors possibles (1, 2, 3 i 4) i elements buits, hi ha", sum(data$cp %in% c
```

```
## [1] "CP: A part dels valors possibles (1, 2, 3 i 4) i elements buits, hi ha 0 registres amb altres va
```

```
if (all.equal(data$trtbps, as.integer(data$trtbps)) ){
  print("trtbps: Tots els valors són enters.")}
```

```
## [1] "trtbps: Tots els valors són enters."
```

```
if (all.equal(data$chol, as.integer(data$chol)) ){
  print("Chol: Tots els valors són enters.")}
```

```
## [1] "Chol: Tots els valors són enters."
```

```
print(paste("fbs: A part dels valors possibles (0, 1), hi ha", sum(data$fbs %in% c(0,1)) - nrow(data) ,
```

```
## [1] "fbs: A part dels valors possibles (0, 1), hi ha 0 registres amb altres valors."
```

```
print(paste("restecg: A part dels valors possibles (0, 1, 2), hi ha", sum(data$restecg %in% c(0,1,2)) -
```

```
## [1] "restecg: A part dels valors possibles (0, 1, 2), hi ha 0 registres amb altres valors."
```

```
if (all.equal(data$thalachh, as.integer(data$thalachh)) ){
  print("thalachh: Tots els valors són enters.")}
```

```
## [1] "thalachh: Tots els valors són enters."
```

```
print(paste("exng: A part dels valors possibles (0, 1), hi ha", sum(data$exng %in% c(0,1)) - nrow(data)
```

```
## [1] "exng: A part dels valors possibles (0, 1), hi ha 0 registres amb altres valors."
```

```
print(paste("slp: A part dels valors possibles (0, 1, 2), hi ha", sum(data$slp %in% c(0,1,2)) - nrow(da
```

```
## [1] "slp: A part dels valors possibles (0, 1, 2), hi ha 0 registres amb altres valors."
```

```
print(paste("caa: A part dels valors possibles (0, 1, 2, 3, 4), hi ha", sum(data$caa %in% c(0,1,2,3,4))
```

```
## [1] "caa: A part dels valors possibles (0, 1, 2, 3, 4), hi ha 0 registres amb altres valors."
```

```
print(paste("thall: A part dels valors possibles (0, 1, 2, 3, 4), hi ha", sum(data$thall %in% c(0,1,2,3
```

```
## [1] "thall: A part dels valors possibles (0, 1, 2, 3, 4), hi ha 0 registres amb altres valors."
```

```
print(paste("output: A part dels valors possibles (0, 1), hi ha", sum(data$output %in% c(0,1,2,3)) - nr
```

```
## [1] "output: A part dels valors possibles (0, 1), hi ha 0 registres amb altres valors."
```

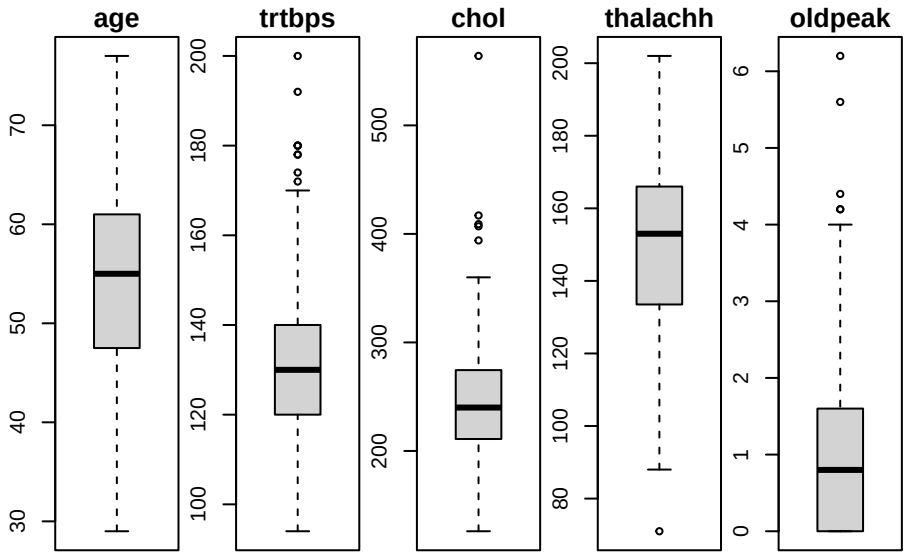
Veiem com la única variable amb anomàlies és "cp" (Tipus de dolor toràcic). Aquesta variable té 143 zeros quan els seus únics valors possibles són 1, 2, 3 o 4. Com es tracta gairebé d'un 50% a zeros, no utilitzarem aquesta variable en la nostra anàlisi, ja que no aportaria informació vàlida.

```
data <- data[, setdiff(names(data), "cp")]
```

3.2. Identifica i gestiona els valors extrems.

Per poder veure les distribucions de les variables i identificar els valors extrems o outliers, presentem uns gràfics de tipus boxplot per a les variables numèriques:

```
data_num <- data[,c('age', 'trtbps', 'chol', 'thalachh', 'oldpeak')]
par(mfrow=c(1,5),mai=c(0.1, 0.0, 0.2, 0.3))
par(oma=c(0, 2, 0, 0) + 0.1)
for (i in colnames(data_num)) boxplot(data_num[i], main = i)
```



En aquests gràfics podem veure que totes les columnes excepte "age" tenen punts individuals fora de les bigues, aquests valors són els que s'han considerat com extrems. A continuació es mostren aquests outliers per a cada variable:

Valors extrems per cada variable amb outliers:

```
res_trtbps <- boxplot(data$trtbps,plot=FALSE)$out
res_chol <- boxplot(data$chol,plot=FALSE)$out
res_thalachh <- boxplot(data$thalachh,plot=FALSE)$out
res_oldpeak <- boxplot(data$oldpeak,plot=FALSE)$out
res_trtbps
```

```
## [1] 172 178 180 180 200 174 192 178 180
```

```
res_chol
```

```
## [1] 417 564 394 407 409
```

```
res_thalachh
```

```
## [1] 71
```

```
res_oldpeak
```

```
## [1] 4.2 6.2 5.6 4.2 4.4
```

Per gestionar aquests valors outliers s'ha decidit passar-los a NA's per imputar-los, després, utilitzant l'algorisme d'aprenentatge no supervisat KNN. Aquest algorisme es basa en el principi que els punts de dades similars tindran valors similars.

Passem a NA tots els outliers:

```
data$trtbps[data$trtbps %in% res_trtbps] <- NA
data$chol[data$chol %in% res_chol] <- NA
data$thalachh[data$thalachh %in% res_thalachh] <- NA
data$oldpeak[data$oldpeak %in% res_oldpeak] <- NA
```

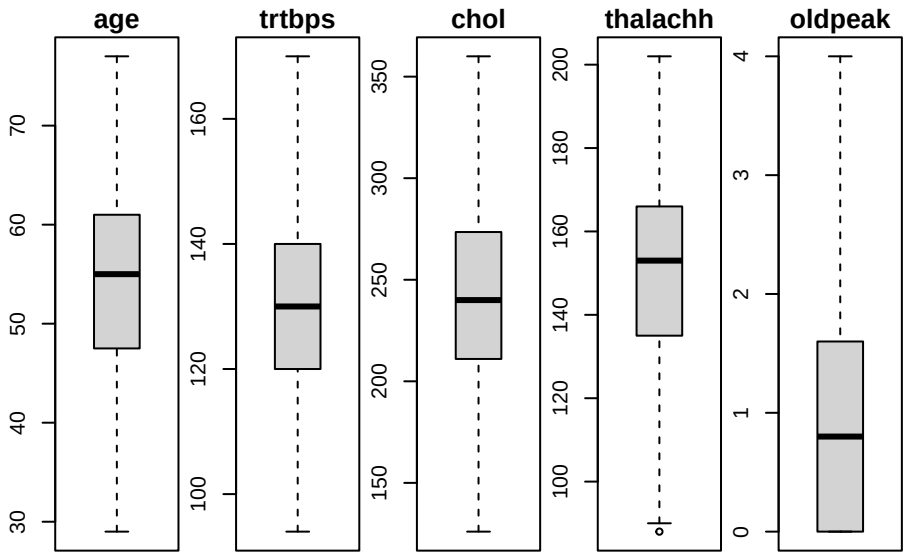
Apliquem KNN a les quatre variables amb un nombre de veïns k igual a 5:

```
if(!require('VIM')) install.packages('VIM'); library('VIM')
dfTrtbpsImp <- kNN(data, variable=c("trtbps"), k = 5)
```

```
dfCholImp <- kNN(data, variable=c("chol"), k = 5)
dfThalachhImp <- kNN(data, variable=c("thalachh"), k = 5)
dfOldpeakImp <- kNN(data, variable=c("oldpeak"), k = 5)
data$trtbps <- dfTrtbpsImp$trtbps
data$chol <- dfCholImp$chol
data$thalachh <- dfThalachhImp$thalachh
data$oldpeak <- dfOldpeakImp$oldpeak
```

Presentem un altre cop el gràfic boxplot per comprovar que efectivament ja no hi ha valors extrems en les dades numèriques:

```
data_num <- data[,c('age', 'trtbps', 'chol', 'thalachh', 'oldpeak')]
par(mfrow=c(1,5),mai=c(0.1, 0.0, 0.2, 0.3))
par(oma=c(0, 2, 0, 0) + 0.1)
for (i in colnames(data_num)) boxplot(data_num[i], main = i)
```



Finalment, es guarda el df en un nou csv

```
write.csv(data, '../Dataset/heart_processed.csv', row.names=TRUE)
```

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar.

A l'hora d'analitzar les dades, pot ser interessant separar les dades en diferents grups:

- En primer lloc, segons el sexe del pacient. Segons el National Heart, Lung, and Blood Institute (NHLBI) estudia desde 2020 les diferències d'atacs cardíacs entre homes i dones joves ja que segons les fonts consultades, com per exemple l'European Society of Cardiology les dones joves que pateixen un atac de cor tenen pitjors resultats que els homes.

Per tant, és interessant separar les dades segons sexe.

```
data_men = data[data$sex == 1,]
data_women = data[data$sex == 0,]
```

- Segons els articles anteriors, podria ser interessant separar també els pacients joves. Els mateixos articles consideren joves a pacients de de 50 anys o menys.

```
data["age_segment"] <- cut(data$age, breaks = c(0,50,120), labels = c(1, 0))
data_young = data[data$age_segment == 1,]
data_old = data[data$age_segment == 0,]
```

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Es comprova la normalitat i la homogeneïtat de la variància de les diferents variables numèriques `age`, `trtbps`, `chol`, `thalachh` i `oldpeak`

- `age`

```
age_t <- data$age
age_m <- data_men$age
age_w <- data_women$age
shapiro.test(age_t); shapiro.test(age_m); shapiro.test(age_w); fligner.test(list(age_m,age_w))
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  age_m
## W = 0.9861, p-value = 0.04
```

```
##
##  Shapiro-Wilk normality test
##
## data:  age_w
## W = 0.97953, p-value = 0.1386
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(age_m, age_w)
## Fligner-Killeen:med chi-squared = 0.517, df = 1, p-value = 0.4721
```

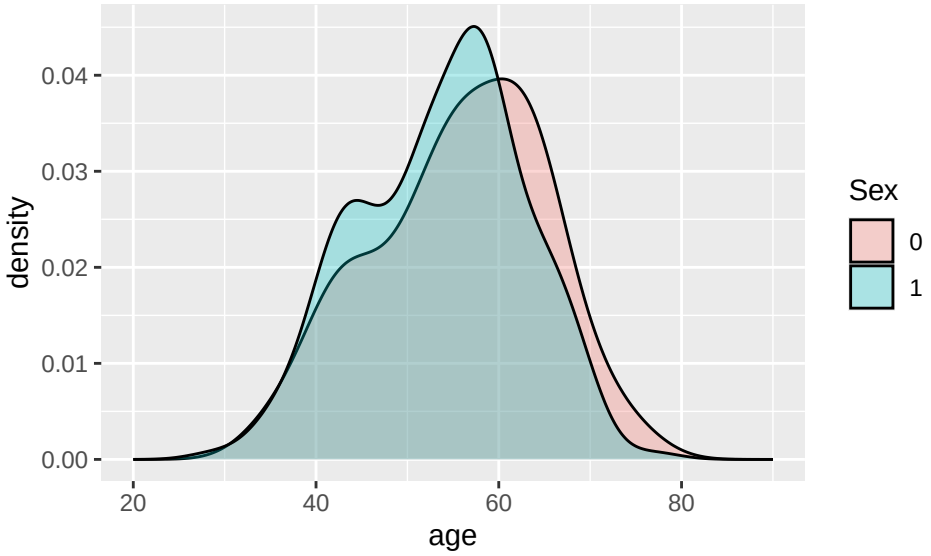
Si es fa un primer anàlisi de la normalitat a tots els pacients en general (`age_t`), el test de normalitat Shapiro-Wilk dóna un `p-value` molt petit i per tant es rebutja la hipòtesis nul · la i no es considera que segueix una distribució normal.

Tot i així, si es separen el pacients entre homes i dones, el test Shapiro-Wilk en el conjunt dels homes té un `p-value` molt proper al nivell de significació generalment utilitzat ($\alpha = 0.05$), i en el cas de les dones és superior. Per tant, es conclou que no es pot rebutjar la hipòtesi nul · la i s'assumeix que les dades segueixen una distribució normal.

El test `fligner.test` de R ens mostra un valor de `p` major a 0.05 i per tant no es pot rebutjar la hipòtesi nul · la d'homoscedasticitat i es conclou que la variable age no presenta variàncies estadísticament diferents per als diferents grups d'homes i dones.

Mostrem les distribucions gràficament:

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
ggplot(data, aes(age, fill = as.factor(sex))) +
  geom_density(alpha = 0.3)+   xlim(20, 90) +
  guides(fill=guide_legend(title="Sex"))
```



- trtbps

```
trtbps_m <- data_men$trtbps
trtbps_w <- data_women$trtbps
shapiro.test(trtbps_m); shapiro.test(trtbps_w); fligner.test(list(trtbps_m,trtbps_w))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trtbps_m
## W = 0.97884, p-value = 0.003279
```

```
##
##  Shapiro-Wilk normality test
##
## data:  trtbps_w
## W = 0.98444, p-value = 0.3161
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(trtbps_m, trtbps_w)
## Fligner-Killeen:med chi-squared = 0.0063353, df = 1, p-value = 0.9366
```

El test Shapiro-Wilk en el conjunt d'homes dona un `p-value` molt petit, però, en canvi, en les dones és més elevat i major al nivell de significació. Per tant, es conclou que les dades del conjunt dels homes no compten amb una distribució normal però el de les dones sí.

Es pot observar a continuació que si es fa una transformació logarítmica ($trtbps\_n = \log(trtbps)$) els `p-value` Shapiro-Wilk test i Fligner-Killeen test són majors que el nivell de significació i, per tant, s'assumeix que les dades segueixen una distribució normal i la variable no presenta variàncies estadísticament diferents per als diferents grups d'homes i dones.

```
data_men$trtbps_n <- log(data_men$trtbps)
data_women$trtbps_n <- data_women$trtbps
shapiro.test(data_men$trtbps_n); shapiro.test(data_women$trtbps_n); fligner.test(list(data_men$trtbps_n
```

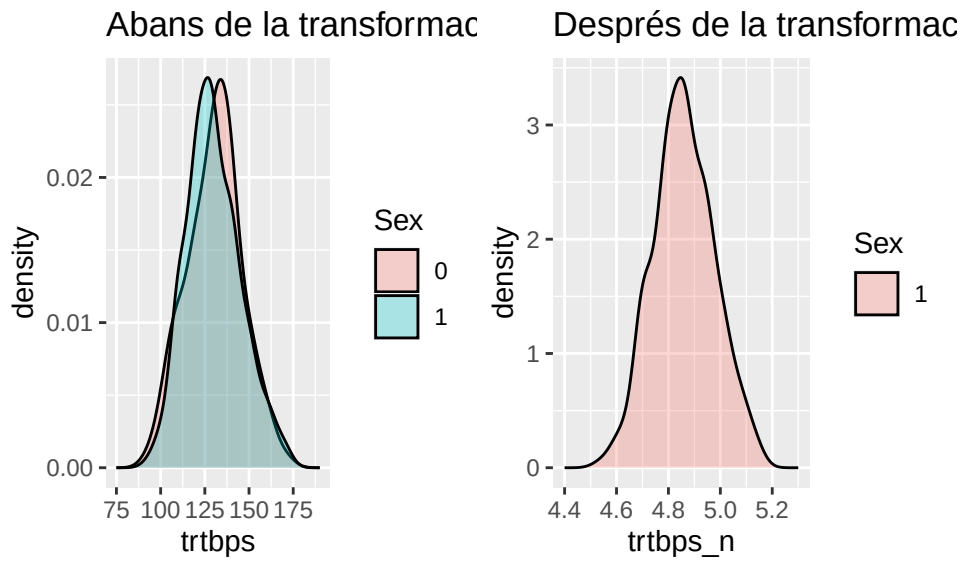
```
##
##  Shapiro-Wilk normality test
##
## data:  data_men$trtbps_n
## W = 0.98708, p-value = 0.0569
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_women$trtbps_n
## W = 0.98444, p-value = 0.3161
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(data_men$trtbps_n, data_women$trtbps_n)
## Fligner-Killeen:med chi-squared = 133.93, df = 1, p-value < 2.2e-16
```

A continuació mostrem gràficament les distribucions de la variable **trtbps** abans, separat en homes i dones, i després de la transformació logarítmica, només homes, ja que la distribució de les dones no ha canviat (ja era normal).

```
if (!require('gridExtra')) install.packages('gridExtra'); library('gridExtra')
p1 <- ggplot(data, aes(trtbps, fill = as.factor(data$sex))) +
  geom_density(alpha = 0.3) + xlim(75, 190) +  guides(fill=guide_legend(title="Sex"))+
  ggtitle(label="Abans de la transformació logarítmica")
p2 <- ggplot(data_men, aes(trtbps_n, fill = as.factor(sex))) +
  geom_density(alpha = 0.3) + xlim(4.4, 5.3) +  guides(fill=guide_legend(title="Sex"))+
  ggtitle(label="Després de la transformació logarítmica")
grid.arrange(p1, p2, nrow=1, ncol=2)
```



- chol

```
chol_m <- data_men$chol
chol_w <- data_women$chol
shapiro.test(chol_m); shapiro.test(chol_w); fligner.test(list(chol_m,chol_w))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  chol_m
## W = 0.99369, p-value = 0.5273
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chol_w
## W = 0.99095, p-value = 0.7647
```

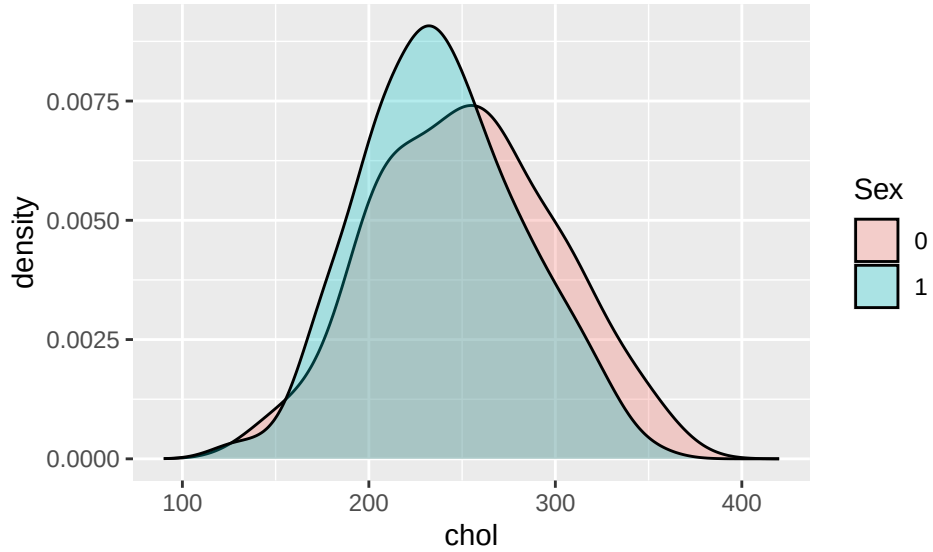
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(chol_m, chol_w)
## Fligner-Killeen:med chi-squared = 2.2055, df = 1, p-value = 0.1375
```

Tant en el cas dels homes com en el de les dones, el `p-value` d'ambdós Shapiro-Wilk test és molt superior al nivell de significació i per tant es conclou que les dades segueixen una distribució normal

El test `fligner.test` de R ens mostra un valor de `p` major a 0.05 i per tant no es pot rebutjar la hipòtesi nul · la d'homoscedasticitat i es conclou que la variable `chol` no presenta variàncies estadísticament diferents per als diferents grups d'homes i dones.

Mostrem les distribucions gràficament:

```
ggplot(data, aes(chol, fill = as.factor(sex))) +
  geom_density(alpha = 0.3)+ xlim(90, 420) +
  guides(fill=guide_legend(title="Sex"))
```



- `thalachh`

```
thalachh_m <- data_men$thalachh
thalachh_w <- data_women$thalachh
shapiro.test(thalachh_m); shapiro.test(thalachh_w); fligner.test(list(thalachh_m,thalachh_w))
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  thalachh_m
## W = 0.98292, p-value = 0.01303
```

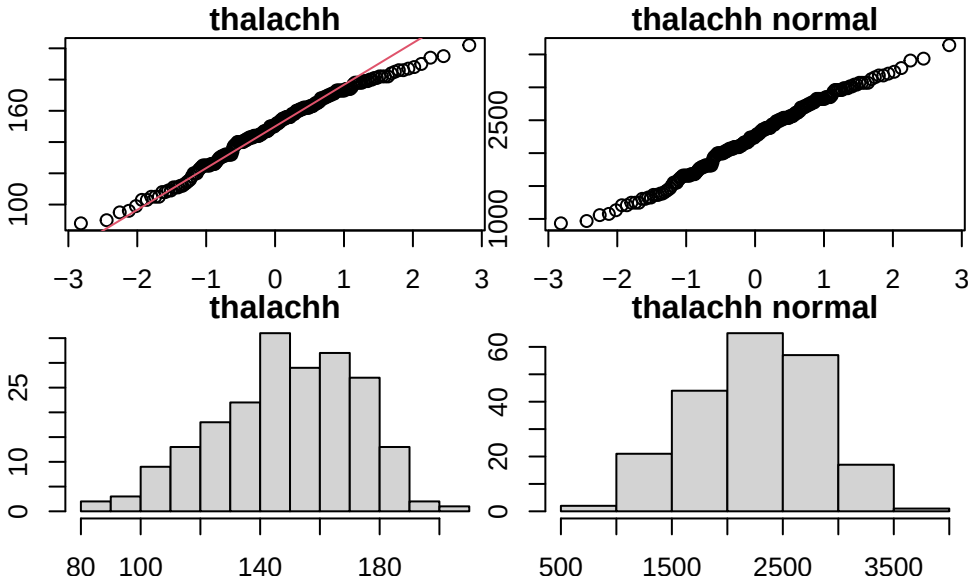
```
##
##  Shapiro-Wilk normality test
##
## data:  thalachh_w
## W = 0.94423, p-value = 0.0004753
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(thalachh_m, thalachh_w)
## Fligner-Killeen:med chi-squared = 4.7508, df = 1, p-value = 0.02928
```

En aquest cas, el test de normalitat Shapiro-Wilk tant el grup d'homes com de dones és molt petit. Així doncs, es conclou que les dades no compten amb una distribució normal. Implementem Box-Cox per normalitzar les distribucions.

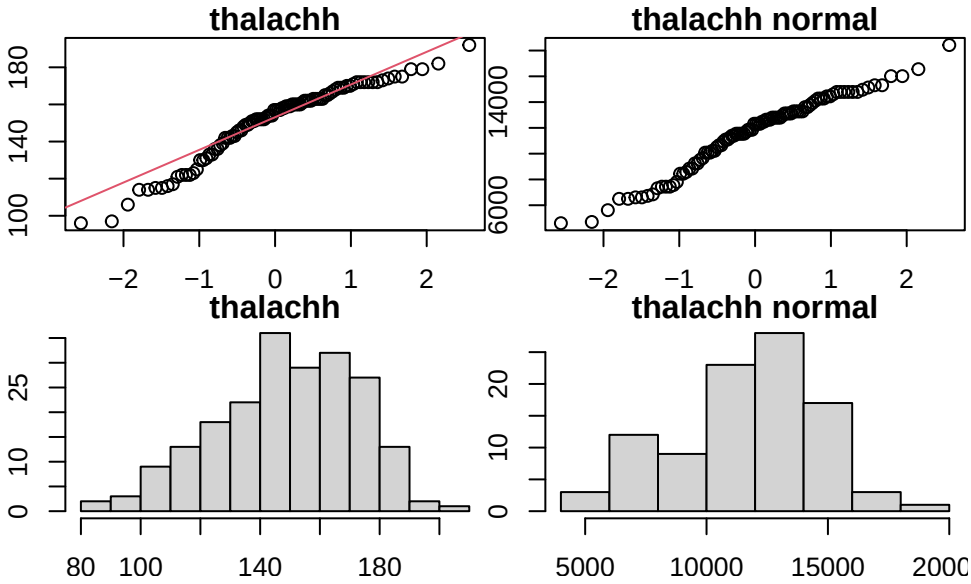
Conjunt d'homes:

```
if (!require('DescTools')) install.packages('DescTools'); library('DescTools')
lambda <- BoxCoxLambda(thalachh_m)
data_men$thalachh_n <- BoxCox(thalachh_m, lambda = lambda)
par(mfrow=c(2,2)); par(mar = c(1.8, 1.8, 1.2, 0.1))
qqnorm(thalachh_m, main="thalachh")
qqline(thalachh_m,col=2)
qqnorm(data_men$thalachh_n, main="thalachh normal")
qqline(thalachh_m,col=2)
hist(thalachh_m,main="thalachh")
hist(data_men$thalachh_n, main="thalachh normal")
```




```
lambda <- BoxCoxLambda(thalachh_w)
data_women$thalachh_n <- BoxCox(thalachh_w, lambda = lambda)
par(mfrow=c(2,2))
par(mar = c(1.8, 1.8, 1.2, 0.1))
qqnorm(thalachh_w, main="thalachh")
qqline(thalachh_w,col=2)
```

```
qqnorm(data_women$thalachh_n, main="thalachh normal")
qqline(thalachh_m,col=2)
hist(thalachh_m,main="thalachh")
hist(data_women$thalachh_n, main="thalachh normal")
```



Tornem a fer el test Shapiro-Wilk per comprovar que les distribucions ja són normals:

```
shapiro.test(data_men$thalachh_n);shapiro.test(data_women$thalachh_n)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_men$thalachh_n
## W = 0.98951, p-value = 0.1352
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data_women$thalachh_n
## W = 0.96737, p-value = 0.01706
```

Veiem com en aquest cas, després de la transformació, el p-value del conjunt dels homes ja és més gran que el nivell de significació i, per tant, es tracta d'una distribució normal. En canvi, pel que fa a les dones, aquesta normalitat, tot i millorar, no s'ha arribat a aconseguir.

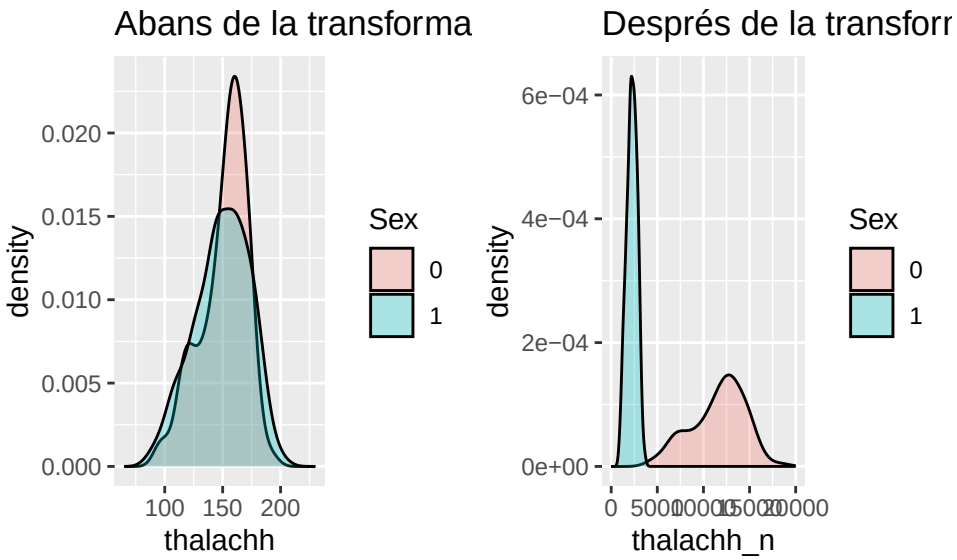
Mostrem les distribucions per sexe gràficament abans i després de la transformació Box-Cox:

```
p1 <- ggplot(data, aes(thalachh, fill = as.factor(sex))) +
  geom_density(alpha = 0.3)+
   xlim(65, 230) +
  guides(fill=guide_legend(title="Sex"))+
  ggtitle(label="Abans de la transformació Box-Cox")

#Es reconstrueix el dataframe amb les dades dels dos gèneres
df <- rbind(data_women, data_men)
df <- df[order(as.numeric(row.names(df))), ]

p2 <- ggplot(df, aes(thalachh_n, fill = as.factor(sex))) +
```

```
  geom_density(alpha = 0.3)+
   xlim(0, 20000) +
  guides(fill=guide_legend(title="Sex"))+
  ggtitle(label="Després de la transformació Box-Cox")
grid.arrange(p1, p2, nrow=1, ncol=2)
```



- oldpeak

```
oldpeak_m <- data_men$oldpeak
oldpeak_w <- data_women$oldpeak
shapiro.test(oldpeak_m)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  oldpeak_m
## W = 0.87552, p-value = 5.125e-12
```

```
shapiro.test(oldpeak_w)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  oldpeak_w
## W = 0.81778, p-value = 1.558e-09
```

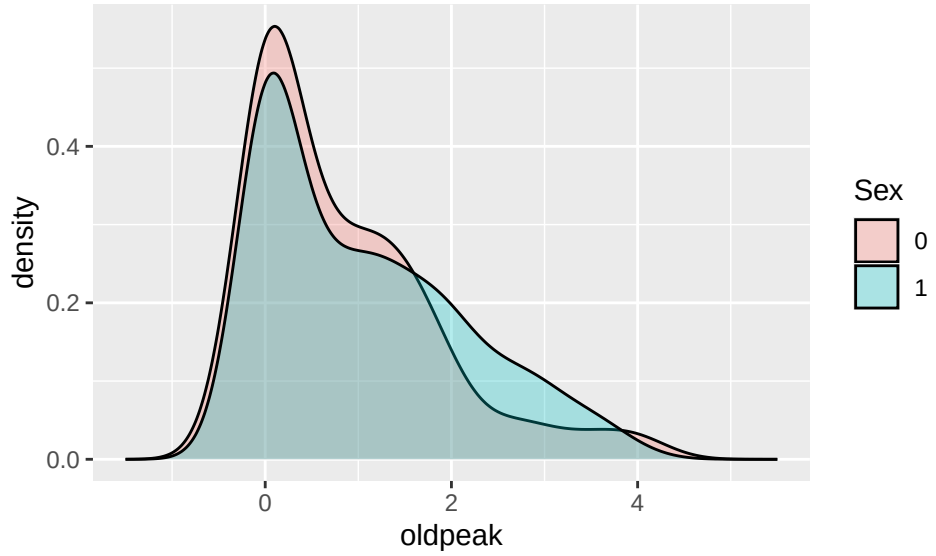
```
fligner.test(list(oldpeak_m,oldpeak_w))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(oldpeak_m, oldpeak_w)
## Fligner-Killeen:med chi-squared = 6.8834, df = 1, p-value = 0.0087
```

Com en el cas anterior, el test de normalitat Shapiro-Wilk tant el grup d'homes com de dones és molt petit. Així doncs, es conclou que les dades no compten amb una distribució normal.

Mostrem les distribucions per sexe gràficament:

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
ggplot(data, aes(oldpeak, fill = as.factor(sex))) +
  geom_density(alpha = 0.3)+
   xlim(-1.5, 5.5) +
  guides(fill=guide_legend(title="Sex"))
```



4.3. Aplicació de proves estadístiques per comparar els grups de dades.

Contrast d'hipòtesis Es tractarà de respondre si:

- L'edat és un factor de risc, tant en homes com en dones.
- Si els nivells de colesterol són un factor de risc, tant en homes com en dones.
- *Edat*: Com la normalitat i l'homoscedasticitat es compleixin (p-valors majors que el nivell de significació), es poden aplicar proves per contrast d'hipòtesis de tipus paramètric, com la prova *t de Student*

La hipòtesis nul · la es que les mitjanes d'edat en homes o dones susceptibles a patir un atac de cor o no són iguals:$H_0 : \mu_{1h/d} = \mu_{0h/d}$

La hipòtesis alternativa és que la mitjanes d'edat en homes o dones susceptibles a patir un atac de cor és superior als que no:$H_1 : \mu_{1h/d} > \mu_{0h/d}$

```
t.test(data_men$age[data_men$output==0],
       data_men$age[data_men$output==1],
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  data_men$age[data_men$output == 0] and data_men$age[data_men$output == 1]
## t = 4.3394, df = 193.88, p-value = 1.149e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.209863      Inf
## sample estimates:
## mean of x mean of y
##  56.08772  50.90323
```

```
t.test(data_women$age[data_women$output==0],
       data_women$age[data_women$output==1],
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  data_women$age[data_women$output == 0] and data_women$age[data_women$output == 1]
## t = 2.8426, df = 81.585, p-value = 0.002826
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.860411      Inf
## sample estimates:
## mean of x mean of y
##  59.04167  54.55556
```

En ambdós casos, el valor `p` obtingut és menor que el nivell de significança i, per tant, es pot rebutjar la hipòtesi nul · la d'igualtat de mitjanes, i com es sospitava, l'edat és un factor de risc.

Per altra banda, es vol comprovar si hi ha diferència entre les mitjanes dels pacients susceptibles a patir un atac de cor entre homes i dones.

La hipòtesis nul · la es que les mitjanes d'edat en homes o dones susceptibles a patir són iguals:$H_0 : \mu_{1h} = \mu_{1d}$

La hipòtesis alternativa és que les mitjanes d'edat en homes o dones susceptibles a patir un atac de cor és superior als que no:$H_1 : \mu_{1h} \neq \mu_{1h}$

```
t.test(data_men$age[data_men$output==1],
       data_women$age[data_men$output==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  data_men$age[data_men$output == 1] and data_women$age[data_men$output == 1]
## t = -3.491, df = 182.43, p-value = 0.0006034
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.304162 -2.029171
## sample estimates:
## mean of x mean of y
##  50.90323  55.56989
```

De nou, s'obté un valor de `p` molt petit, per tant es pot rebutjar la hipòtesis nul · la i es conclou que la mitja d'edat en homes és diferent a la de dones.

- *Nivell de colesterol*: Com la normalitat i l'homoscedasticitat es compleixin (p-valors majors que el nivell de significació), es poden aplicar proves per contrast d'hipòtesis de tipus paramètric, com la prova *t de Student*.

Es consideren les mateixes hipòtesis que l'apartat anterior, però en aquest cas en comptes de l'edat es considera el nivell de colesterol.

```
t.test(data_men$chol[data_men$output==0],
       data_men$chol[data_men$output==1],
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  data_men$chol[data_men$output == 0] and data_men$chol[data_men$output == 1]
## t = 2.6021, df = 204.89, p-value = 0.004971
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5.501347      Inf
## sample estimates:
## mean of x mean of y
##  246.0614  230.9892
```

```
t.test(data_women$chol[data_women$output==0],
       data_women$chol[data_women$output==1],
       alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  data_women$chol[data_women$output == 0] and data_women$chol[data_women$output == 1]
## t = 1.3361, df = 42.243, p-value = 0.09433
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.757751       Inf
## sample estimates:
## mean of x mean of y
##  263.2917  248.7639
```

És interessant veure que, segons el contrast d'hipòtesis, s'obtenen resultats diferents en homes i dones. En el cas dels homes el valor p obtingut és menor que el nivell de significança i, per tant, es pot rebutjar la hipòtesi nul · la i es conclou que el nivell de colesterol és un factor de risc en homes. Per altra banda, en dones surt un valor de p més elevat i no es pot rebutjar la hipòtesis nul · la.

- **thalachh**: Com la normalitat i l'homoscedasticitat no es compleixen no es poden aplicar proves per contrat d'hipòtesis de tipus paramètric però es poden aplicar proves no paramètriques com Wilcoxon.

```
wilcox.test(data_men$thalachh, data_women$thalachh, alternative="greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  data_men$thalachh and data_women$thalachh
## W = 9468.5, p-value = 0.7453
## alternative hypothesis: true location shift is greater than 0
```

El resultat del test NO permet rebutjar la hipòtesi nul · la d'igualtat de freqüència cardíaca màxima aconseguida i, per tant, es conclou que la freqüència cardíaca màxima aconseguida no presenta diferències estadísticament significatives entre homes i dones.

- Proporció d'infarts entre homes i dones.

La hipòtesis nul · la és que les proporcions d'atac al cor entre homes i dones són les mateixes:$H_0 : p_h = p_d$

La hipòtesis alternativa és que la proporció d'atacs al cor és superior en homes que en dones.:$H_1 : p_h > p_d$

```
prop.test(x=c(dim(data_women[data_women$output==1,])[1],
              dim(data_men[data_men$output==1,])[1]),
          n = c(dim(data_women)[1],
                dim(data_men)[1]),
          alternative= "greater",
          conf.level=0.95,
          correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(dim(data_women[data_women$output == 1, ])[1], dim(data_men[data_men$output == 1, ])[1]) out
## X-squared = 23.914, df = 1, p-value = 5.036e-07
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2084305 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7500000 0.4492754
```

El `p-valor` és inferior a 0.05, per tant, es pot concloure que, segons les dades proporcionades, la proporció d'atacs al cor és superior en homes que en dones.

Correlació Es representen les correlacions entre parells de variables continues en un conjunt de dades:

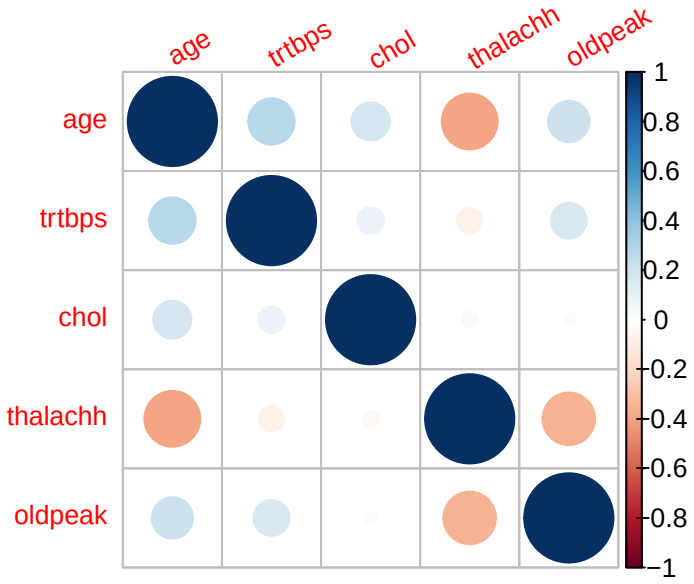
```
tau_corr <- cor(data_num)
print(tau_corr)
```

```
##                 age      trtbps        chol    thalachh     oldpeak
## age       1.0000000  0.27077921  0.18275660 -0.39228187  0.21484156
## trtbps    0.2707792  1.00000000  0.08853152 -0.07974415  0.16190269
## chol      0.1827566  0.08853152  1.00000000 -0.03251327  0.02068918
## thalachh -0.3922819 -0.07974415 -0.03251327  1.00000000 -0.34783002
## oldpeak   0.2148416  0.16190269  0.02068918 -0.34783002  1.00000000
```

Es visualitza gràficament:

```
if(!require("corrplot")) install.packages("corrplot"); library("corrplot")
if(!require('dplyr')) install.packages('dplyr'); library('dplyr')

res <- cor(data_num)
corrplot(res,
         upper="number",
         tl.srt=30,
         number.cex=0.7,
         tl.cex=0.8)
```



Es pot observar que els atributs amb més correlació són:

- `age` (edat) i `thalachh` (Pressió arterial en repòs) ~40%: Generalment, els valors normals de la pressió aterial en repós augmenten amb l'edat.
- `thalachh` (Pressió arterial en repòs) i `oldpeak` (Depressió de ST induïda per l'exercici en relació amb el descans) ~35%: La Depressió de ST es refereix a una troballa en un electrocardiograma , en què el traç en el segment ST és anormalment baix per sota de la línia de base. Això pot ser degut a certes dolències (sistèmiques o no) que també afecten a la pressió arterial.
- La resta, tenen correlacions inferiors al 35%.

En tots els casos, es considera que no hi ha una correlació prou significativa (>|50%|) com per eliminar l'atribut.

Regressió logística A continuació utilitzarem una regressió logística per intentar predir si una persona té o no més d'un 50% de possibilitats de patir malalties cardíaques. Aquest tipus d'anàlisi de regressió s'utilitza per predir variables binàries a partir d'altres variables. En el nostre cas utilitzarem totes les dades seleccionades per predir la variable output.

A R, aquest tipus de models s'estimen mitjançant la funció `glm()`, especificant la família com a binomial.

Primer, factoritzem les variables categòriques i divideix el dataset "desordenat" en 2/3, que s'utilitzaran per l'entrenament, i el 1/3 restant, per testejar.

```
data$sex <- as.factor(data$sex)
data$fbs <- as.factor(data$fbs)
data$restecg <- as.factor(data$restecg)
data$exng <- as.factor(data$exng)
data$slp <- as.factor(data$slp)
data$caa <- as.factor(data$caa)
data$thall <- as.factor(data$thall)
data$output <- as.factor(data$output)
```

```
set.seed(1234)
data_shuffled <- data[sample(nrow(data)),]
rows = 2/3 * nrow(data)
data_train <- data_shuffled[1: rows, ]
data_test <- data_shuffled[rows:nrow(data), ]
```

```
dim(data); dim(data_train); dim(data_test)
```

```
## [1] 303  14
```

```
## [1] 202  14
```

```
## [1] 102  14
```

A continuació executem la regressió i mostrem un resum dels resultats:

```
bc <- data[complete.cases(data_train),]
logit <- glm(output ~. ,data=bc, family="binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = output ~ ., family = "binomial", data = bc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7294  -0.4064   0.1610   0.4742   2.9191
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.152335   3.364342   0.640  0.52234
## age          -0.013415   0.037641  -0.356  0.72155
## sex1         -1.427340   0.489303  -2.917  0.00353 **
## trtbps       -0.015978   0.012423  -1.286  0.19838
## chol         -0.009241   0.004689  -1.971  0.04876 *
## fbs1          0.770024   0.531745   1.448  0.14759
## restecg1      0.398720   0.376723   1.058  0.28988
## restecg2     -0.854446   1.720513  -0.497  0.61945
## thalachh      0.018121   0.011074   1.636  0.10178
## exng1        -1.329928   0.420717  -3.161  0.00157 **
## oldpeak      -0.356817   0.229314  -1.556  0.11970
## slp1         -0.385408   0.680711  -0.566  0.57127
## slp2          0.746236   0.759408   0.983  0.32578
## caa1         -2.056177   0.490073  -4.196 2.72e-05 ***
## caa2         -3.209613   0.694060  -4.624 3.76e-06 ***
## caa3         -2.490459   0.895776  -2.780  0.00543 **
## caa4          0.752433   1.574156   0.478  0.63266
## thall1        2.292738   1.848774   1.240  0.21492
## thall2        2.809827   1.749436   1.606  0.10824
## thall3        1.029560   1.736377   0.593  0.55322
## age_segment0  0.874579   0.713586   1.226  0.22034
## ---
## 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: 199.64  on 282  degrees of freedom
## AIC: 241.64
##
## Number of Fisher Scoring iterations: 6
```

La bondat del model s'avalua mitjançant la mesura AIC (criteri d'informaicó d'Akike), en el nostre model aquest valor és 241.64. Com més petit és aquest valor, millor és la bondat de l'ajustament i menor la complexitat del model.

També veiem com el nombre de vasos principals del cor (variable caa) és molt significatiu per al model, seguit de exng (exercici angina de pit induïda), el sexe i, finalment, el colesterol.

A continuació testejem el model amb el dataset reservat i mostrem la matriu de confusió corresponent.

```
predict <- predict(logit, data_test, type = 'response')
# confusion matrix
conf_mat <- table(data_test$output, predict > 0.5)
conf_mat
```

```
##
##     FALSE TRUE
##   0    36    8
##   1     5   53
```

Descrivim els resultats de la matriu de confusió:

- True Negative (TN): 36 casos en els quals el model ha predit correctament que no hi ha un infart.
- False Positive (FP): 8 casos en els quals el model ha predit un infart però no n'hi havia.
- False Negative (FN): 5 casos en els quals el model ha predit que no hi havia un infart però en realitat n'hi havia.
- True Positive (TP): 53 casos en els quals el model ha predit correctament un infart.

A partir d'aquests valors podem extreure mètriques per avaluar l'efectivitat del model en funció del nostre objectiu. En el nostre cas d'estudi, és interessant maximitzar el nombre de positius vertaders, ja que es vol assegurar que el màxim nombre de casos de malalties cardíaques es detectin correctament i, per tant, ser capaços de predir persones amb alta probabilitat de patir un atac de cor.

Per això, és particularment important obtenir un model amb una sensibilitat alta. Igualment, aquesta alta sensibilitat, no hauria de venir acompanyada d'una baixa especificitat, que significaria un gran nombre de falsos positius, és a dir, persones que sense patir malaltia cardíaca es detecten com a malalts.

A continuació mostrem els resultats de les mètriques:

```
recall <- conf_mat[2,2] / sum(conf_mat[2,])
accuracy <- sum(diag(conf_mat)) / sum(conf_mat)
especif <- conf_mat[1,1] / sum(conf_mat[1,])
recall
```

```
## [1] 0.9137931
```

```
accuracy
```

```
## [1] 0.872549
```

```
especif
```

```
## [1] 0.8181818
```

Podem observar com la sensibilitat, o recall, té un valor alt i, per tant, el model entra dintre de les necessitats del nostre problema. L'accuracy i l'especificitat també són altes, però menys, especialment l'especificitat, fet que indica que potser hi ha més falsos positius dels desitjats.

5. Resolució del problema.

El problema inicialment plantejat és si aquest dataset, "Heart Attack Analysis & Prediction", pot determinar quins són els factors de risc més importants per a la salut cardiovascular.

Es pot concloure que:

- Mitjançant el contrast d'hiòtesis:
  - En general, l'edat és un factor de risc. La mitja d'edat de pacients amb probabilitat de patir un atac de cor `output=1` és més elevada als que no.
  - Per altra banda, s'observa que les mitjanes d'edat entre homes i dones amb probabilitat de patir un atac de cor és diferent.
  - Pel que fa als nivells de colesterol, s'ha observat que el nivell de colesterol és un factor de risc en homes mentre que en el cas de les dones no es pot descartar la hipòtesis nul · la.
  - Segons les dades proporcionades, la proporció d'atacs al cor és superior en homes que en dones
- Correlació líneal: tant `age` (edat) i `thalachh` (Pressió arterial en repòs) com `thalachh` (Pressió arterial en repòs) i `oldpeak` (Depressió de ST induïda per l'exercici en relació amb el descans) ténen una correlació d'entre el 35% i 40%. La resta, tenen correlacions inferiors al 35%. En tots els casos, es considera que no hi ha una correlació prou significativa (>|50%|) com per eliminar l'atribut.
- Regressió logistica: S'ha entrenat i validat un model de regressió logística utilitzant totes les dades seleccionades després de la neteja per predir la variable output (probabilitat de patir malalties del cor).

  El model ha demostrat una alta sensibilitat, del 91%, indicant que és capaç de detectar correctament la majoria dels casos positius. Així mateix, l'accuracy i l'especificitat també són alts, del 87% i 81% respectivament. Això indica que el model és generalment precís en la seva predicció. No obstant, l'especificitat és lleugerament baixa, cosa que significa que hi ha un nombre relativament alt de falsos positius.

Finalment, concloure que, en la línia del nostre objectiu, s'ha extret un anàlisis rellevant sobre el conjunt de les dades i la seva relació amb la probabilitat de patir malalties cardíaques. Tot i així, es destaca que el dataset utilitzat no disposa de gaire volum de dades i, per tant, es considera que aquests resultats podrien no ser del tot fiables. Seria recomenable ampliar el dataset i repetir l'estudi per disminuir l'error i obtenir una major confiança en les conclusions extretes.

6. Vídeo.

https://drive.google.com/drive/u/1/folders/1FxrpxE3qg98JGHOX872uEIdfW40zAeFU

7. Taula de contribucions.

Contribucions	Signatura
Investigació prèvia	Judith Cid i Jordi Tormo
Redacció de les respostes	Judith Cid i Jordi Tormo
Desenvolupament del codi	Judith Cid i Jordi Tormo
Participació al vídeo	Judith Cid i Jordi Tormo