

Tipologia i cicle de vida de les dades: Pràctica 2

Autors: Àngels Calvet i Mirabent i Albert Estadella Valls

Juny 2022

Contents

Apartat 1	2
Apartat 2	3
Apartat 3	4
Apartat 3.1.	4
Apartat 3.2.	5
Apartat 3.3.	9
Apartat 4	10
Apartat 4.1.	11
Apartat 4.2.	11
Apartat 4.3.	11
Apartat 6	15
Taula contribucions	16

Instal·lem i carreguem les llibreries necessàries per a la realització de la pràctica.

```
# https://cran.r-project.org/web/packages/nortest/index.html
if (!require('nortest')) install.packages('nortest'); library('nortest')
# https://cran.r-project.org/web/packages/ResourceSelection/index.html
if (!require('ResourceSelection')) install.packages('ResourceSelection');
library('ResourceSelection')
```

Apartat 1

Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset utilitzat va ser publicat a la web de Kaggle (<https://www.kaggle.com/datasets>) per l'usuari 'NARESH BHAT' l'any 2020 sota el nom 'Health care: Heart attack possibility'. Consisteix en una selecció d'observacions i atributs d'una base de dades més gran. Aquest dataset es pot trobar amb la següent URL: '<https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility>'

El dataset conté extreptes de la base de dades de Cleveland Clinic Foundation amb l'objectiu de poder determinar si el pacient té més o menys risc de patir un atac de cor.

En concret, el dataset conté 303 observacions corresponents a diferents pacients i les següents 14 variables:

- **age**: Edat pacient
- **sex**: Sexe pacient
 - 0: Dona
 - 1: Home
- **cp**: Tipus de dolor de pit
 - 0: Angina típica
 - 1: Angina atípica
 - 2: Dolor no angina
 - 3: Asimptomàtic
- **trestbps**: pressió arterial en repòs (en mm Hg)
- **chol**: Colesterol en mg/dl
- **fbs**: Fasting blood sugar > 120 mg/dl
 - 0: Fals -> Pacient no diabètic
 - 1: Cert -> Pacient diabètic
- **restecg**: Resultat electrocardiograma
 - 0: Normal
 - 1: Anormalitat en el segment ST de l'electrocardiograma
 - 2: Hipertrofia ventricular esquerra
- **thalach**: Màxim ritme cardíac registrat
- **exang**: Angina induïda per exercici
 - 0: No
 - 1: Sí
- **oldpeak**: Depressió del segment ST induïda per exercici relatiu al descans.
- **slope**: Pendent del segment ST en el pic d'exercici
 - 0: Positiu
 - 1: Pla
 - 2: Negatiu
- **ca**: Número de grans vasos (0-3) colorejants amb fluoroscopia
- **thal**: talassèmia (existència d'un desordre en el flux sanguini)
 - 1: Flux sanguini normal
 - 2: Defectes fixes (irreversibles, no trobem flux en alguna part del cor)
 - 3: Defectes reversibles (flux observat però no normal)
- **target**: variable que indica si el pacient té probabilitat de patir un atac de cor
 - 0: menor probabilitat de patir atac de cor

- 1: major probabilitat de patir atac de cor

Les malalties cardiovasculars com pot ser l'atac de cor són una de les principals causes de mortalitat en els països desenvolupats. Aquest fet fa que el dataset sigui rellevant donat que permetrà analitzar i determinar quins són els factors, mesurats en àmbit clínic, que tenen un major impacte en determinar si el pacient ha patit o no un atac de cor.

A més a més de poder determinar quins són els factors amb major influència, per una banda es generarà una regressió que permeti decidir si un pacient té risc de patir un atac de cor, i per altra banda, a partir de contrastos d'hipòtesi es buscarà donar respostes a preguntes tals com:

- Tenen el mateix risc de patir un atac de cor els homes i les dones?
- Hi ha una relació significativa entre el colesterol i la possibilitat de patir un atac de cor?

Apartat 2

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.

Carreguem el joc de dades.

```
# Càrrega del fitxer
df <- read.csv('heart.csv', stringsAsFactors = FALSE)
# Visualització primers registres del dataset
head(df)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  63  1  3   145   233   1       0    150     0     2.3    0  0    1
## 2  37  1  2   130   250   0       1    187     0     3.5    0  0    2
## 3  41  0  1   130   204   0       0    172     0     1.4    2  0    2
## 4  56  1  1   120   236   0       1    178     0     0.8    2  0    2
## 5  57  0  0   120   354   0       1    163     1     0.6    2  0    2
## 6  57  1  0   140   192   0       1    148     0     0.4    1  0    1
##   target
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

Observem com tots els atributs poden ser útils per a determinar el risc de patir un atac de cor, per tant, es decideix seguir l'anàlisi amb tots ells. Observem també que algunes variables nominals estan definides com a tipus enter, per tant, s'han de convertir a tipus factor.

```
# Preparació de les dades
df$sex <- as.factor(df$sex)
df$cp <- as.factor(df$cp)
df$sex <- as.factor(df$sex)
df$fbs <- as.factor(df$fbs)
df$restecg <- as.factor(df$restecg)
df$exang <- as.factor(df$exang)
df$slope <- as.factor(df$slope)
df$thal <- as.factor(df$thal)
df$target <- as.factor(df$target)
head(df)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  63  1  3    145  233   1      0    150    0    2.3    0  0    1
## 2  37  1  2    130  250   0      1    187    0    3.5    0  0    2
## 3  41  0  1    130  204   0      0    172    0    1.4    2  0    2
## 4  56  1  1    120  236   0      1    178    0    0.8    2  0    2
## 5  57  0  0    120  354   0      1    163    1    0.6    2  0    2
## 6  57  1  0    140  192   0      1    148    0    0.4    1  0    1
##   target
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

Apartat 3

Neteja de les dades.

Apartat 3.1.

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

- Comprovem si hi ha nuls i/o valors buits.

```
# Nombre total de valors buits i nuls
colSums(is.na(df))
```

```
##      age      sex      cp trestbps      chol      fbs restecg  thalach
##      0        0        0         0         0        0         0         0
##  exang  oldpeak  slope      ca      thal  target
##      0        0        0         0         0         0
```

```
colSums(df=="")
```

```
##      age      sex      cp trestbps      chol      fbs restecg  thalach
##      0        0        0         0         0        0         0         0
##  exang  oldpeak  slope      ca      thal  target
##      0        0        0         0         0         0
```

Observem com no hi ha valors nuls o buits a tractar.

- Revisem si hi ha zeros en les variables numèriques.

Disseminem les variables categòriques i numèriques.

```
numVar <- c("age", "trestbps", "chol", "thalach", "oldpeak")
catVar <- c("sex", "cp", "fbs", "restecg", "exang", "slope", "ca", "thal", "target")
```

Comprovem si hi ha zeros

```
# Nombre total de valors iguals a 0
colSums(df[numVar]==0)
```

```
##      age trestbps      chol  thalach  oldpeak
##      0         0         0         0        99
```

Únicament la variable 'oldpeak' té registres amb valor zero, però com aquest poden ser valors possibles, no es realitza cap tractament addicional.

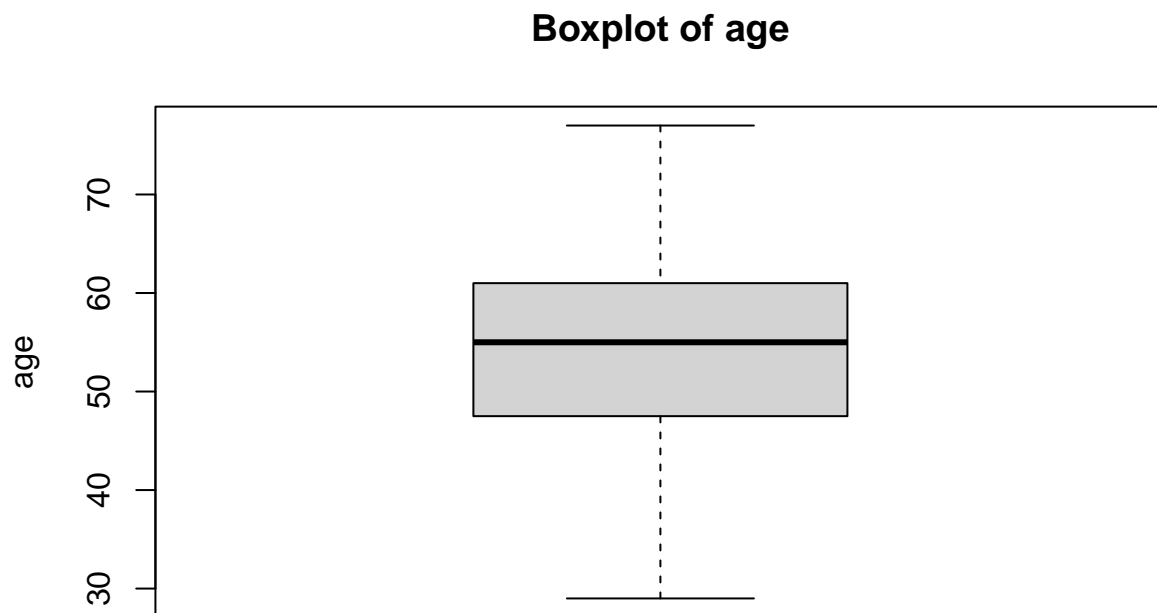
Apartat 3.2.

Identifica i gestiona els valors extrems.

Per a la detecció de valors atípics dins les variables numèriques, es fa ús dels diagrames de caixes, amb ells podrem detectar visualment els possibles valors anòmals i de la funció `boxplots.stats()`, que retorna els valors anòmals.

Variable 'age'

```
# Boxplot
boxplot(df$age, main='Boxplot of age', ylab= 'age')
```



```
# Get outliers
boxplot.stats(df$age)$out
```

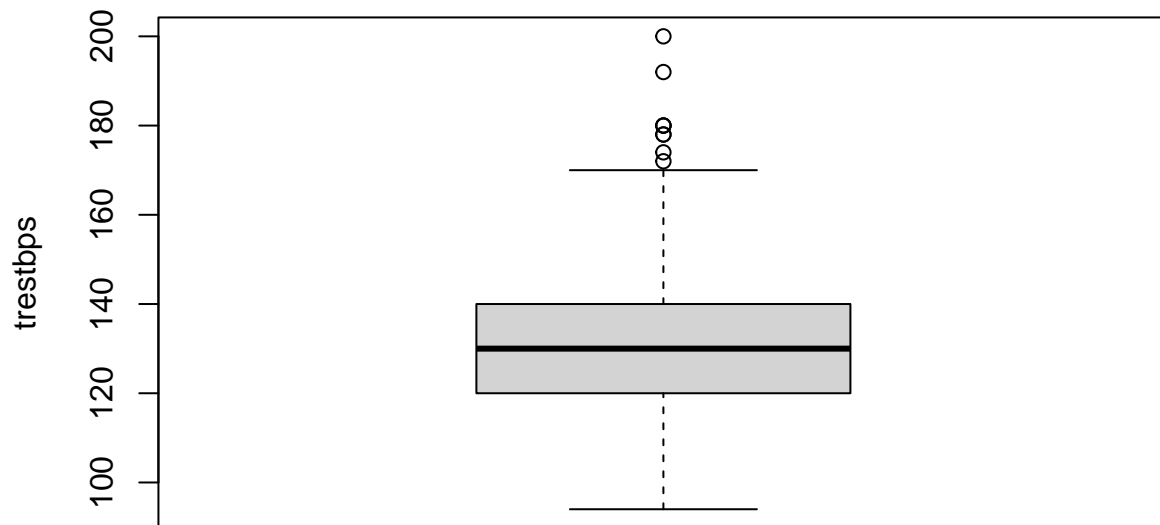
```
## integer(0)
```

No es detecten valors anòmals per cap dels dos mètodes.

Variable 'trestbps'

```
# Boxplot
boxplot(df$trestbps, main='Boxplot of resting blood pressure', ylab= 'trestbps')
```

Boxplot of resting blood pressure



```
# Get outliers  
boxplot.stats(df$trestbps)$out
```

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

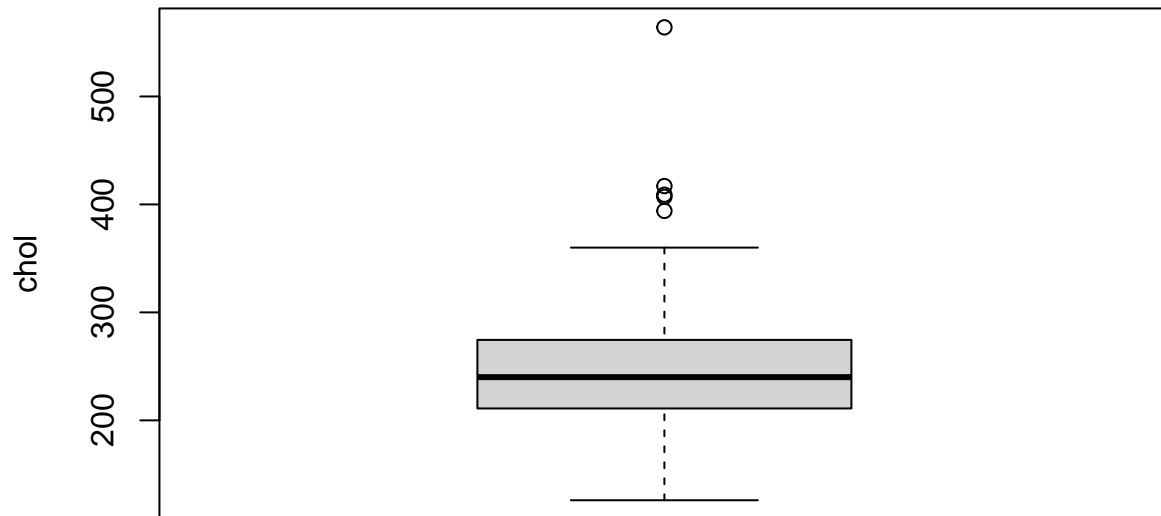
Es detecten valors anòmals tant gràficament com numèricament.

Es considera pressió arterial normal aquella pressió per sota els 120 mmHg. Valors superiors als 180 mmHg són possibles d'assolir en casos de crisis d'hipertensió, per tant, els outliers detectats no s'eliminaran ja que donat que han estat obtinguts en pacients ingressats a l'hospital poden ser perfectament factibles. Fins i tot, fa pensar que podria ser un dels factors determinants de patir un atac de cor.

Variable 'chol'

```
# Boxplot  
boxplot(df$chol, main='Boxplot of cholesterol', ylab= 'chol')
```

Boxplot of cholesterol



```
# Get outliers
boxplot.stats(df$chol)$out
```

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

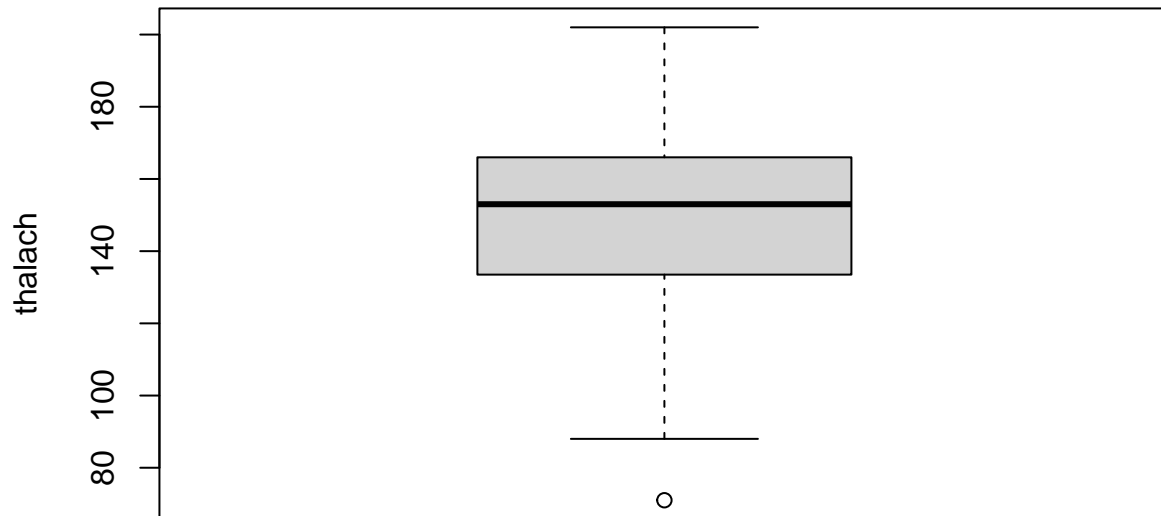
Es detecten valors anòmals tant gràficament com numèricament.

Valors de colesterol per sobre dels 240 mg/dl, es consideren elevat i ja poden tenir impactes negatius sobre la salut. Mesures per sobre els 500 mg/dl són rares, però possibles, per tant, no es realitzarà cap tractament sobre els registres amb aquests valors.

Variable 'thalach'

```
# Boxplot
boxplot(df$thalach, main='Boxplot of maximum heart rate achieved', ylab= 'thalach')
```

Boxplot of maximum heart rate achieved



```
# Get outliers
boxplot.stats(df$thalach)$out
```

```
## $stats
## [1] 88.0 133.5 153.0 166.0 202.0
##
## $n
## [1] 303
##
## $conf
## [1] 150.05 155.95
##
## $out
## [1] 71
```

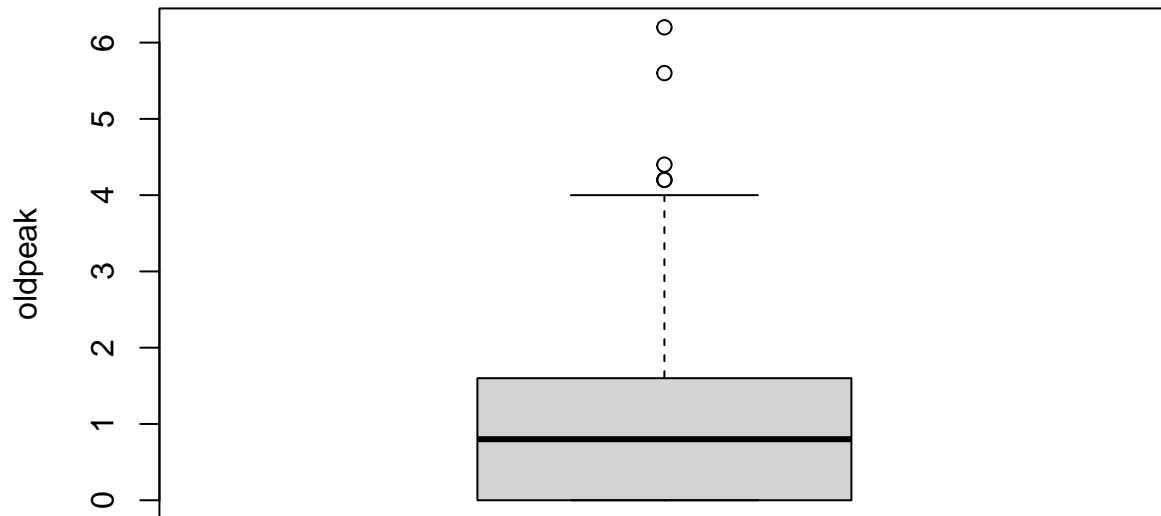
Es detecta una única observació anòmla amb un valor de 71. La freqüència cardíaca màxima és el nombre de pulsacions per minut a les que el cor és capaç de bombejar a màxima pressió. Tenint en compte que la forma d'estimar aquest paràmetre és restant 220 menys la teva edat, el valor de 71 és realment molt baix, ja que en persones d'uns 80 anys (rang superior que trobem d'edat), estariem parlant de valors de 140 (el doble de 70). Per tant, es decideix eliminar els outliers.

```
# Eliminació registres amb freqüència cardíaca inferior a 71
dfClean <- df[df$thalach > 71, ]
```

Variable 'oldpeak'

```
# Boxplot
boxplot(df$oldpeak, main='Boxplot of ST depression induced by exercise relative to rest', ylab= 'oldpeak')
```


Boxplot of ST depression induced by exercise relative to rest



```
# Get outliers
boxplot.stats(df$oldpeak)$out
```

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

Es detecten valors anòmals tant gràficament com numèricament. El segment ST representa el període isoelectric, és a dir, quan els ventricles es troben entre la despolarització i la repolarització. Per tant, no hi ha flux elèctric conseqüentment hauria de ser pla aproximadament a la línia basal. El segment ST deprimit és aquell fenomen en què la línia del segment es troba per sota de la basal i la qual sabem que pot estar associada a un infart de miocardi. La depressió es mesura en mm per sota el complex QRS. Una depressió superior a 1mm ja és significativa. Per tant, els valors anòmals que trobem són realment elevats considerant els rangs. Tot i això, no els eliminarem ja que una possibilitat futura seria convertir aquesta variable numèrica es una variable categòrica on els valors normals es trobessin entre 0 i 1, depressió entre 1 i 2 i alta depressió valors >2.

Apartat 3.3.

Respecte a les variables categòriques, únicament cal comprovar que les etiquetes corresponen amb les que indica el creador del dataset.

```
# Possibles valors variables categòriques
sapply(dfClean[catVar], table)
```

```
## $sex
##
##  0  1
## 96 206
##
```

```
## $cp
##
## 0 1 2 3
## 142 50 87 23
##
## $fbs
##
## 0 1
## 257 45
##
## $restecg
##
## 0 1 2
## 147 151 4
##
## $exang
##
## 0 1
## 203 99
##
## $slope
##
## 0 1 2
## 21 139 142
##
## $ca
##
## 0 1 2 3 4
## 174 65 38 20 5
##
## $thal
##
## 0 1 2 3
## 2 18 165 117
##
## $target
##
## 0 1
## 137 165
```

Observem com la variable 'thal' té dos registres amb un valor de "0", valor inexistent en els definits pel creador del dataset, per tant, procedim a eliminar aquests 2 registres.

```
# Eliminació registres amb valor 0 de la variable 'thal'
dfClean <- dfClean[!dfClean$thal=="0", ]
```

Finalment, es guarda el nou dataset en un nou fitxer de nom Heart_Clean.csv:

```
# Exportació de les dades netes en .csv
write.csv(dfClean, "Heart_Clean.csv")
```

Apartat 4

Anàlisi de les dades.

Apartat 4.1.

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?).

Se seleccionen grups dins el conjunt de dades que poden resultar interessants per analitzar o comparar.

```
# Agrupació per diagnosi (probabilitat de patir atac de cor)
dfClean.atac.no <- dfClean[dfClean$target == "0",] # Probabilitat baixa
dfClean.atac.si <- dfClean[dfClean$target == "1",] # Probabilitat alta
```

Apartat 4.2.

Comprovació de la normalitat i homogeneïtat de la variància.

- Normalitat

Amb el test de 'Shaphiro' i de 'Lilliefors', podem comprovar la normalitat d'una variable. En aquest cas, com disposem de més de 300 observacions, es procedeix a realitzar el test 'Lilliefors'.

```
for (i in numVar) {
  pvalue <- lillie.test(dfClean[,i])$p.value # Test Lilliefors
  print(paste0('P valor test Lilliefors per atribut ', i, ': ', pvalue ))
}
```

```
## [1] "P valor test Lilliefors per atribut age: 0.000129167916497149"
## [1] "P valor test Lilliefors per atribut trestbps: 6.04352991582741e-08"
## [1] "P valor test Lilliefors per atribut chol: 0.0380179349936097"
## [1] "P valor test Lilliefors per atribut thalach: 0.00151017070902159"
## [1] "P valor test Lilliefors per atribut oldpeak: 4.34602899434539e-28"
```

Es considera que es segueix una distribució normal en cas que el valor de la probabilitat (p-value) sigui major a 0.05. Com podem observar, cap variable numèrica del dataset d'estudi segueix una distribució normal.

- Homogeneïtat de la variància

```
# test homoscedasticitat
var.test(dfClean.atac.no$chol,dfClean.atac.si$chol)
```

```
##
## F test to compare two variances
##
## data: dfClean.atac.no$chol and dfClean.atac.si$chol
## F = 0.85515, num df = 135, denom df = 163, p-value = 0.3466
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.6201425 1.1850999
## sample estimates:
## ratio of variances
## 0.8551456
```

El test considera com hipòtesi nul · la que les variàncies són iguals, per tant, en obtenir un p valor superior a 0.05, es conclou que les variàncies són iguals. En aquest cas, no es pot rebutjar hipòtesi nul · la ($p > 0.05$), per tant, sí que hi ha homoscedasticitat.

Apartat 4.3.

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.

- **Correlació**

Es realitza un anàlisi de correlació per determinar si existeix correlació entre les diferents variables numèriques i quines són les que influeixen més sobre la probabilitat de patir un atac de cor.

```
# Càlcul correlació
cor(dfClean[numVar], method = 'spearman')

##           age    trestbps      chol    thalach    oldpeak
## age      1.0000000  0.29137849  0.19523298 -0.39566392  0.26789182
## trestbps 0.2913785  1.00000000  0.12540249 -0.04707741  0.15540363
## chol     0.1952330  0.12540249  1.00000000 -0.05086240  0.04367986
## thalach  -0.3956639 -0.04707741 -0.05086240  1.00000000 -0.44233696
## oldpeak  0.2678918  0.15540363  0.04367986 -0.44233696  1.00000000
```

No observem una correlació significativa entre les diferents variables numèriques.

Respecte a la variable objectiu 'target', les màximes correlacions s'obtenen amb els atributs 'thalach' i 'oldpeak' tenint aquest uns valors d'aproximadament 0.4, valors que indiquen que no hi ha molta correlació. Per tant, podríem concloure que no existeix una relació directa entre les diferents variables numèriques.

- **Pregunta 1: Les variables sexe i target estan relacionades o són independents?**

Per respondre a la pregunta es realitza un test de contrast d'hipòtesi amb les següents hipòtesis:

Hipòtesi nul · la (H_0): les variables sexe i target són independents.

Hipòtesi alternativa (H_1): les variables sexe i target són dependents, existeix una relació entre elles.

```
# Matriu contingència (valors observats)
tableCont <- table(dfClean$sex, dfClean$target)
row.names(tableCont) <- c("F","M")
print("Matriu de contingència")

## [1] "Matriu de contingència"

print(tableCont)

##
##      0    1
## F   24   71
## M  112   93

tableCont2 <- prop.table(tableCont, margin = 2)
print("Matriu de contingència amb percentatges respecte la probabilitat de tenir un atac de cor")

## [1] "Matriu de contingència amb percentatges respecte la probabilitat de tenir un atac de cor"

print(tableCont2)

##
##      0      1
## F 0.1764706 0.4329268
## M 0.8235294 0.5670732

tableCont3 <- prop.table(tableCont, margin = 1)
print("Matriu de contingència amb percentatges respecte el sexe")

## [1] "Matriu de contingència amb percentatges respecte el sexe"

print(tableCont3)
```

```
##
##           0           1
##  F 0.2526316 0.7473684
##  M 0.5463415 0.4536585
```

Per respondre la pregunta, el primer pas a seguir ha sigut el càlcul de la matriu de contingència. La primera matriu ens mostra el nombre de pacients amb més (1) o menys (0) probabilitat de patir un atac de cor en funció del sexe; i la segona i tercera matrius són la mateixa però en percentatge (una respecte la probabilitat de patir un atac de cor i l'altre respecte el sexe). Podem observar que el nombre d'homes de l'estudi en general (més de 200) és molt més elevat que el de dones (no arriba a 100). Per tant, els resultats no seran equitatius. Una altra observació important és la proporció de pacients amb probabilitat d'atac de cor. De tota la mostra observada gairebé el 75% de les dones la probabilitat de patir un atac de cor és alta, en canvi, en els homes és aproximadament del 50%.

```
# chisq.test
chisq.test(tableCont)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tableCont
## X-squared = 21.427, df = 1, p-value = 3.675e-06
```

Donat que el p valor és inferior al nivell de significança 0.05, pràcticament és 0, podem rebutjar la hipòtesi nul·la i concloure que per un nivell de confiança del 95%, hi ha una relació de dependència entre les variables 'sex' i 'target'.

- **Pregunta 2: El colesterol dels pacients que no han patit un atac de cor és significativament inferior al dels pacients que sí han patit atac de cor?**

Per respondre a la pregunta es realitza un test de contrast d'hipòtesi amb les següents hipòtesis:

Hipòtesi nul·la (H_0): $\mu_{chol0} - \mu_{chol1} = 0$

Hipòtesi alternativa (H_1): $\mu_{chol0} - \mu_{chol1} < 0$

Mitjana colesterol per a observacions amb variable 'target' = 0: μ_{chol0}

Mitjana colesterol per a observacions amb variable 'target' = 1: μ_{chol1}

```
t.test(dfClean.atac.no$chol, dfClean.atac.si$chol, var.equal=TRUE, alternative = "less",
       conf.level=0.95)
```

```
##
## Two Sample t-test
##
## data:  dfClean.atac.no$chol and dfClean.atac.si$chol
## t = 1.52, df = 298, p-value = 0.9352
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 19.07542
## sample estimates:
## mean of x mean of y
## 251.5368 242.3902
```

El p valor obtingut és molt proper a 1 i no podem rebutjar la hipòtesi nul·la. Per tant, no podem afirmar que tenir un colesterol elevat és motiu d'augment de la probabilitat de patir un atac de cor.

A l'hora d'interpretar aquest resultat, s'ha de tenir en compte que tal com podem veure en el boxplot de la variable colesterol, la gran major part dels pacients del dataset, tenen un nivell de colesterol elevat, poden així fer disminuir la incidència del colesterol sobre la possibilitat de patir un atac de cor.

- Regressió logística

Per poder predir la probabilitat de patir un atac de cor s'ajustarà un model de regressió logística amb totes les variables disponibles, ja que per literatura i els resultats sabem que d'alguna manera hipotetitzem que influeix en la probabilitat de patir un atac de cor.

```
# Create model
model.logist <- glm(formula = 'target~age+sex+cp+trestbps+chol+fbs+restecg+thalach+exang+
                        oldpeak+slope+ca+thal',
                    family=binomial(link=logit), data=dfClean, na.action=na.omit)

# Model information
summary(model.logist)
```

```
##
## Call:
## glm(formula = "target~age+sex+cp+trestbps+chol+fbs+restecg+thalach+exang+\n
##      family = binomial(link = logit), data = dfClean, na.action = na.omit)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7867  -0.3619   0.1584   0.5303   2.5622
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.9106997  2.7835351   1.046  0.29571
## age          0.0004444  0.0235226   0.019  0.98493
## sex1        -1.4152336  0.5303272  -2.669  0.00762 **
## cp1          0.9573668  0.5639733   1.698  0.08959 .
## cp2          1.8841075  0.4799999   3.925 8.66e-05 ***
## cp3          1.9570504  0.6514233   3.004  0.00266 **
## trestbps     -0.0169515  0.0106985  -1.584  0.11309
## chol         -0.0041501  0.0038778  -1.070  0.28452
## fbs1          0.2135194  0.5682434   0.376  0.70710
## restecg1      0.6169639  0.3781427   1.632  0.10277
## restecg2     -0.2733710  2.2731570  -0.120  0.90428
## thalach       0.0155424  0.0113049   1.375  0.16918
## exang1       -0.7826272  0.4289293  -1.825  0.06806 .
## oldpeak      -0.4894948  0.2258116  -2.168  0.03018 *
## slope1       -0.6977302  0.8589508  -0.812  0.41662
## slope2        0.1830203  0.9343980   0.196  0.84471
## ca           -0.8396432  0.2038834  -4.118 3.82e-05 ***
## thal2         0.1383618  0.7775826   0.178  0.85877
## thal3        -1.2937485  0.7559739  -1.711  0.08701 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 413.27  on 299  degrees of freedom
## Residual deviance: 200.55  on 281  degrees of freedom
## AIC: 238.55
##
## Number of Fisher Scoring iterations: 6
```

Hem de tenir en compte que en totes les variables categòriques tenim un coeficient per nivell-1, ja que aquests

es calculen respecte el primer: $\text{sex} = 0$, $\text{cp} = 0$, $\text{fbs} = 0$, $\text{restecg} = 0$, $\text{exang} = 0$, $\text{slope} = 0$ i $\text{thal} = 0$. El resultat de la regressió logística dependrà dels coeficients multiplicats pel valor numèric de les variables numèriques i més els coeficients corresponents a les variables categòriques d'aquell cas. Observem les variables sex , cp , oldpeak i ca són significatives, és a dir, que tenen si un d'ells varia, l'efecte en el resultat pot variar significativament. D'aquestes, podem veure que el sexe té un efecte negatiu respecte el sexe femení, és a dir, que quan el sexe és masculí (1) el resultat disminuirà i per tant, la probabilitat de patir un atac de cor també, per tant, concorda amb els resultats trobats amb la matriu de contingència. La variable cp té un efecte positiu en tots els casos respecte a tenir un tipus de dolor d'angina típica. I les variables oldpeak i ca també tenen un efecte negatiu, respectivament s'interpretaria com: si augmenta la depressió del segment ST, la probabilitat de patir un atac de cor disminuirà; si augmenta el nombre de grans vasos colorejats amb fluoroscopia, la probabilitat de patir un atac de cor disminuirà, fet que té sentit ja que si no els veiem podria ser degut a una obstrucció.

Variables com l'edat, les quals podríem pensar que haurien de tenir un efecte significatiu en el resultat podria ser a causa de la mostra escollida. Hem de tenir en compte, com podem veure en el boxplot anterior, que la mostra analitzada està formada majoritàriament per persones d'entre 47 i 61 anys, per tant, el model no s'ajustaria tant bé fora d'aquest rang.

Per a determinar la bondat de l'ajust fem servir el test Hosmer-Lemeshow.

```
hoslem.test(model.logist$y, fitted(model.logist))
```

```
##  
## Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data: model.logist$y, fitted(model.logist)  
## X-squared = 4.3453, df = 8, p-value = 0.8247
```

Amb el test de Hosmer-Lemeshow obtenim un p-valor superior a 0.05, resultat que indica que no s'ha de rebutjar la hipòtesi nul·la (no hi ha diferències entre els valors observats i els pronosticats), per tant, es pot determinar que el model està ben ajustat.

Apartat 6





Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Amb els resultats obtinguts podem concloure alguns fets, tot i això, el problema no el podem considerar resolt. L'objectiu del projecte consistia en, per una banda, determinar quins són els factors amb major influència en que una persona pateixi un atac de cor, i per l'altra banda, donar resposta a dues preguntes:

- Tenen el mateix risc de patir un atac de cor els homes i les dones?
- Hi ha una relació significativa entre el colesterol i la possibilitat de patir un atac de cor?

Pel primer objectiu podem concloure que les variables que tenen un major efecte segons els nostres resultats són el sexe, la depressió del segment ST i el nombre de grans vasos colorejats amb fluoroscopia. Tot i això, hem de tenir en compte que la mostra en termes de sexe no està equilibrada, i que les edats que alberga són una mica limitades. Tot i això, basant-nos en els resultats podríem dir que les dones tenen una tendència major a patir atacs de cor i que no existeix una relació significativa entre el colesterol i la possibilitat de patir un atac de cor, tot i que no considerem que siguin uns resultats fiables.

Taula contribucions

Contribucions	Signatura	
<i>Investigació prèvia</i>		
<i>Redacció de les respostes</i>		
<i>Desenvolupament del codi</i>	