title “Pràctica_2”

Authors

“Roger Alvarez” “Xavier Borrat”

date: “2023-01-06” output: pdf_document

LINKS a les fonts de les dades:

https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

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

El dataset de treball es una modificació d'un dataset original del que hem pogut treure més informació per entendre millor les variables. (https://archive.ics.uci.edu/ml/datasets/Heart+Disease)

El dataset de treball conté informació sobre pacients que consulten sobre diferent simptomatologia de probable origen cardíac. Aquesta informació es el resultat de registrar la resposta a diferents preguntes i proves per determinar si els simptomes corresponen a obstrucció coronària verificada per una angiografia coronoria. L'angiografia o cateterisme es la prova “gold standard” per a determinar si un pacient està patint o té un alt risc de patir un infart de cor.

Al dataset hi ha diferents variables que descriuen diferents aspectes dels pacients, com ara l'edat(age), el gènere(sex), si han experimentat dolor de pit durant l'exercici físic(exang), el nombre de vasos sanguinis calcificats en una radiografia continua(escopia)(ca), el tipus de dolor de pit que han experimentat(cp), la pressió arterial en repòs(trtbps), la quantitat de colesterol en sang(chol), si han fet una prova de tolerància a la glucosa en dejú(fbs), el resultat de l'electrocardiograma en repòs(rest_ecg) i la freqüència cardíaca màxima aconseguida durant la prova d'esforç(thalach). Totes aquestes variables actuarien com a potencials predictors de risc de presentar o no obstrucció de les artèries coronàries.

Finalment hi ha una variable diana que indica si el pacient té més o menys probabilitats de patir un atac de cor mesurat a través de l'obtrucció dels vasos coronaris a través d'un catetisme cardíac.

```
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ggplot2)
setwd("~/Documents/GitHub/datathon2021/heart_attack")
# Carrega  de l'arxiu de dades

heart <- read.csv("heart.csv")
```

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.

```
# 2. Integracció i selecció
# Seleccionem les columnes que volem:

heart_sel<- heart[c("age", "sex", "trtbps", "chol", "fbs", "restecg", "exng", "caa", "output")]

# Convertim a tipus factor les variables discretes.

#sex: sex (1 = male; 0 = female)
heart_sel$sex<-factor(heart_sel$sex)

#fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
heart_sel$fbs<-factor(heart_sel$fbs)

#restecg: Lectures de Electrocardiograma  en repòs.
#Value 0: normal
#Value 1: having ST-T wave abnormality
heart_sel$restecg<-factor(heart_sel$restecg)


heart_sel$exng<-factor(heart_sel$exng)


# output Value 0: < 50% diameter narrowing
# output Value 1: > 50% diameter narrowing
#heart_sel$output<-factor(heart_sel$output)

#Funció resum del dataset.
summary(heart_sel)
```

```
##       age        sex         trtbps           chol       fbs     restecg
##  Min.   :29.00   0: 96   Min.   : 94.0   Min.   :126.0   0:258   0:147
##  1st Qu.:47.50   1:207   1st Qu.:120.0   1st Qu.:211.0   1: 45   1:152
##  Median :55.00           Median :130.0   Median :240.0           2:  4
##  Mean   :54.37           Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00           3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00           Max.   :200.0   Max.   :564.0
##  exng         caa             output
##  0:204   Min.   :0.0000   Min.   :0.0000
##  1: 99   1st Qu.:0.0000   1st Qu.:0.0000
##          Median :0.0000   Median :1.0000
##          Mean   :0.7294   Mean   :0.5446
##          3rd Qu.:1.0000   3rd Qu.:1.0000
##          Max.   :4.0000   Max.   :1.0000
```

3. Neteja de les dades.

```
# 3.1 0 Rastreig de  valors. En aquest cas no hi ha valors buits.

colSums(is.na(heart_sel))
```

```
##     age     sex  trtbps    chol     fbs restecg    exng     caa  output
##       0       0       0       0       0       0       0       0       0
```

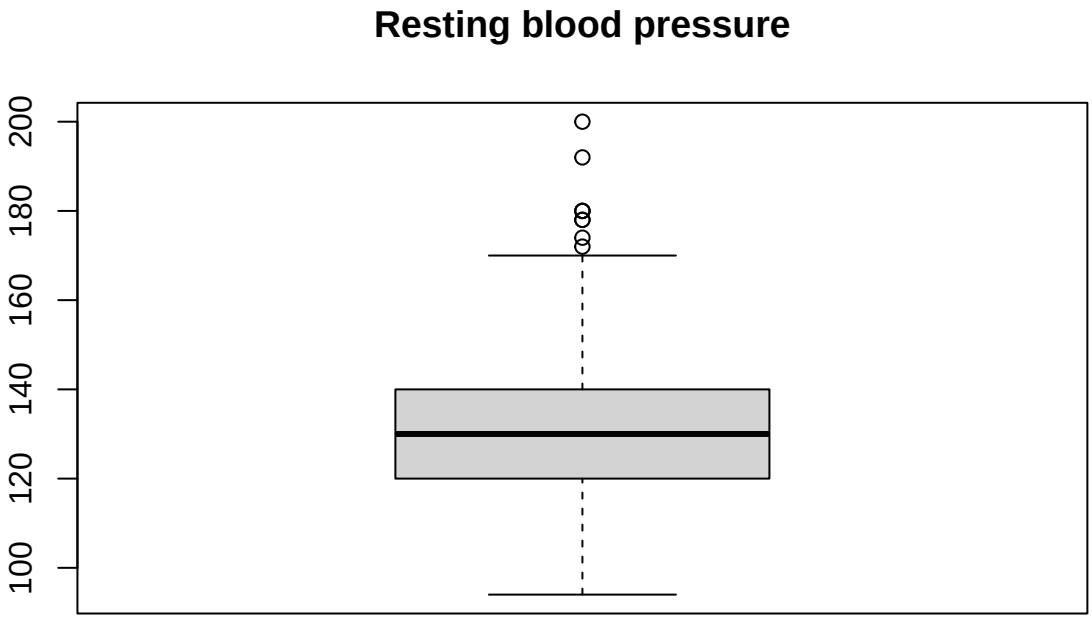
Transformem la veriable restecg on passem de tres valors de la variable a nominal a només 2. El motiu és que clínicament un registre d'hipertrofia en el context de la malaltia coronària comptabilitazaria com a normal per tant el valor 2 el convertirem a valor 0.

```
# En el cas de la variable restecg la convertim en dicotòmica convertint els valors 2 en 0.

heart_sel$restecg[heart_sel$restecg == 2] <- 0
heart_sel$restecg<-factor(heart_sel$restecg)
table(heart_sel$restecg)
```

```
##
##   0   1
## 151 152
```

Indentificació de valors extrems(outlayer) utilitzarem com a definició una desviació de més de 1.5 vegades el rang interquantilic per sobre el quartil superior i per sota el quartil inferior (Q1 - 1.5 * IQR or Q3 + 1.5 * IQR).

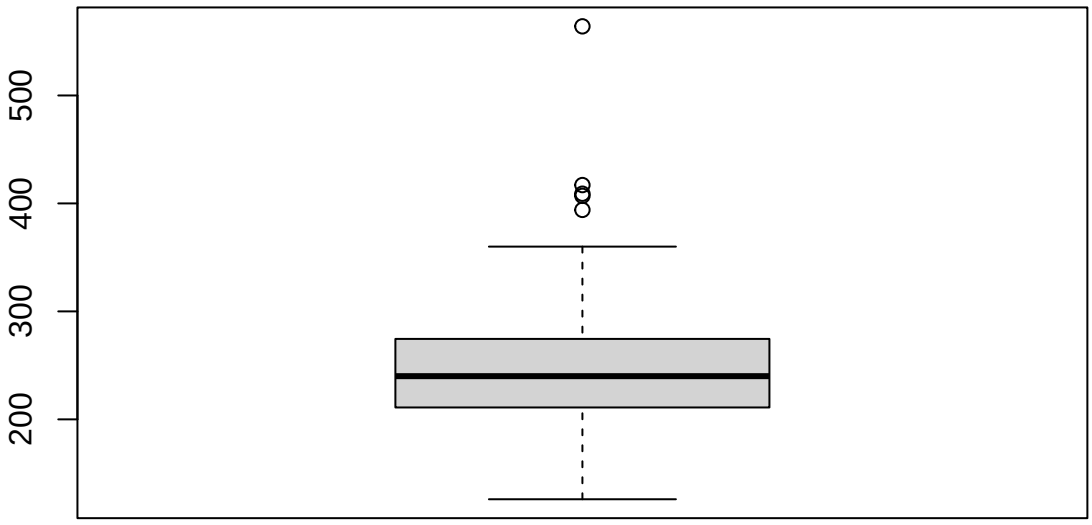
```
# 3.2 valors extrems


boxplot(heart_sel$trtbps, main="Resting blood pressure")
```




```
rbp_outliers <- boxplot.stats(heart_sel$trtbps)$out
sort(rbp_outliers)
```

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

```
boxplot(heart_sel$chol, main="Colesterol level")
```




```
col_outliers <- boxplot.stats(heart_sel$chol)$out
sort(col_outliers)
```

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

```
# Encara que alguns dels valors  siguin plausibles i altres no, ens agafarem a la definicó estadísitca

heart_sel$trtbps[heart_sel$trtbps >= min(rbp_outliers)] <- NA
heart_sel$chol[heart_sel$chol >= min(col_outliers)] <- NA


# Imputem valors nous utilitzan KNN.

heart_sel<- kNN(heart_sel, variable= "trtbps", k = 11)
heart_sel<- kNN(heart_sel, variable= "chol", k = 11)
```

4. Anàlisi de les dades.

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?). **Objectiu principal**:

Amb aquest conjunt de dades volem generar un model predictiu per estimar el diagnòstic de malaltia coronaria i veure la importancia relativa en el seu diagnòstic de dades biomètriques, analítiques i de tests d'estrés.

Regressio logística per predir la malaltia coronaria utilitzant com a variables independents: -Edat(Age) -Sexe(Sex) -Glicèmia basal com a marcador de diabetis.(fps) -Colesterol en sang.(chol) -Pressió arterial en repòs.(trtbpm) -Dolor en repòs.(exng) -Alteració segment ST en repòs.(restecg) -Nombre d'arteries coronaries calcificades. (output)

Objectiu Secundari

Abans però estudiarem el comportament d'algunes variables per entendre millor el comporatament de les dades. Entre elles:

1.-Relació entre dolor anginós en exercici (exng) i presentar elevació del segment ST(restecg) com a marcador d'isquèmia miocàrdica. Matriu de confusió: exng, restecg

```
# Relació entre dolor anginós a l'exercici (exng) i presentar elevació del segment ST(restecg) com a ma

#H0: The two variables are independent.
#H1: The two variables relate to each other.

chisq.test(heart_sel$exng, heart_sel$restecg)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  heart_sel$exng and heart_sel$restecg
## X-squared = 2.2797, df = 1, p-value = 0.1311
```

Conclusió: No hi ha relació entre la ocorrència d'aquests dos fets i per tant son indendents. Tot i que el dolor precordial en repòs és altament suggestiu de coronariopatia, no s'associa a elevació del segment ST en repós ja que aquesta mesura en aquest dataset s'obté en el context d'una prova d'esforç que es una prova per diagnosticar angines d'esforç i no angines de repòs. Per tant quan hi ha dolor en repós la prova diagnótica a fer no es una prova d'esforç per això es plausible que en aquest context les variables pughin ser independents.

2.-Quin és el nivell de relació entre: -edat(age) vs colesterol(chol) -edat(age) vs. pressió arterial en repòs(trtbps).

Correlació: edat vs colesterol edat vs pressió arterial en repòs

```
# Quin és el nivell de relació entre edat(age) i colesterol(chol) i edat i pressió arterial en repós(tr

# Comprovem la normalitat de les distribucions de les dades mitjançant el test shapiro-wilkins i l'apro

# The null-hypothesis of this test is that the population is normally distributed.

shapiro.test(heart_sel$age)
```

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

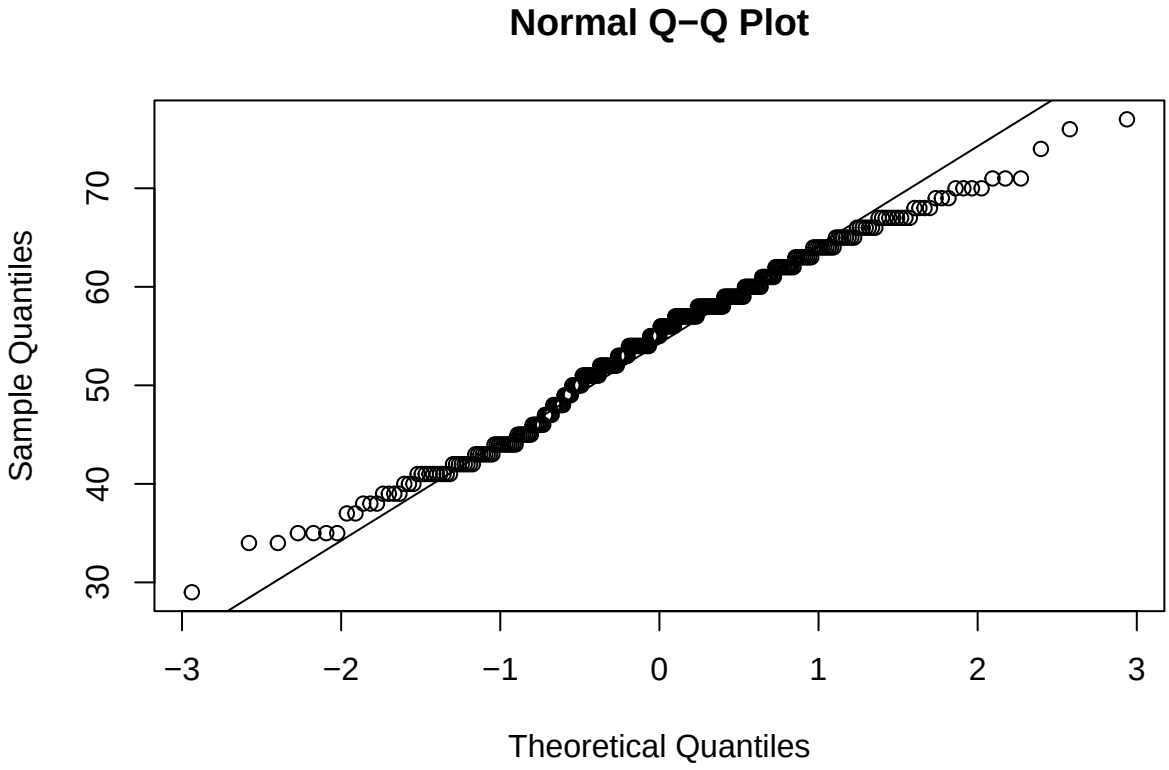
```
shapiro.test(heart_sel$chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heart_sel$chol
## W = 0.99404, p-value = 0.2801
```

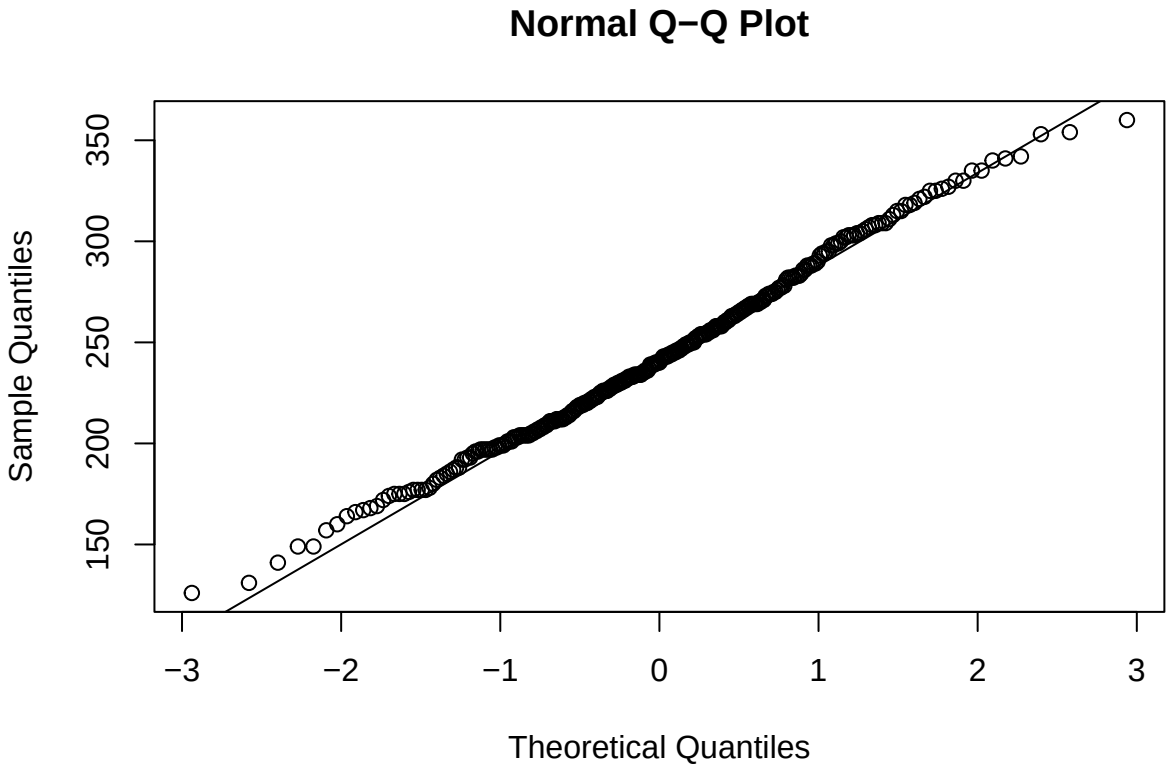
```
shapiro.test(heart_sel$trtbps)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  heart_sel$trtbps
## W = 0.98514, p-value = 0.003193
```

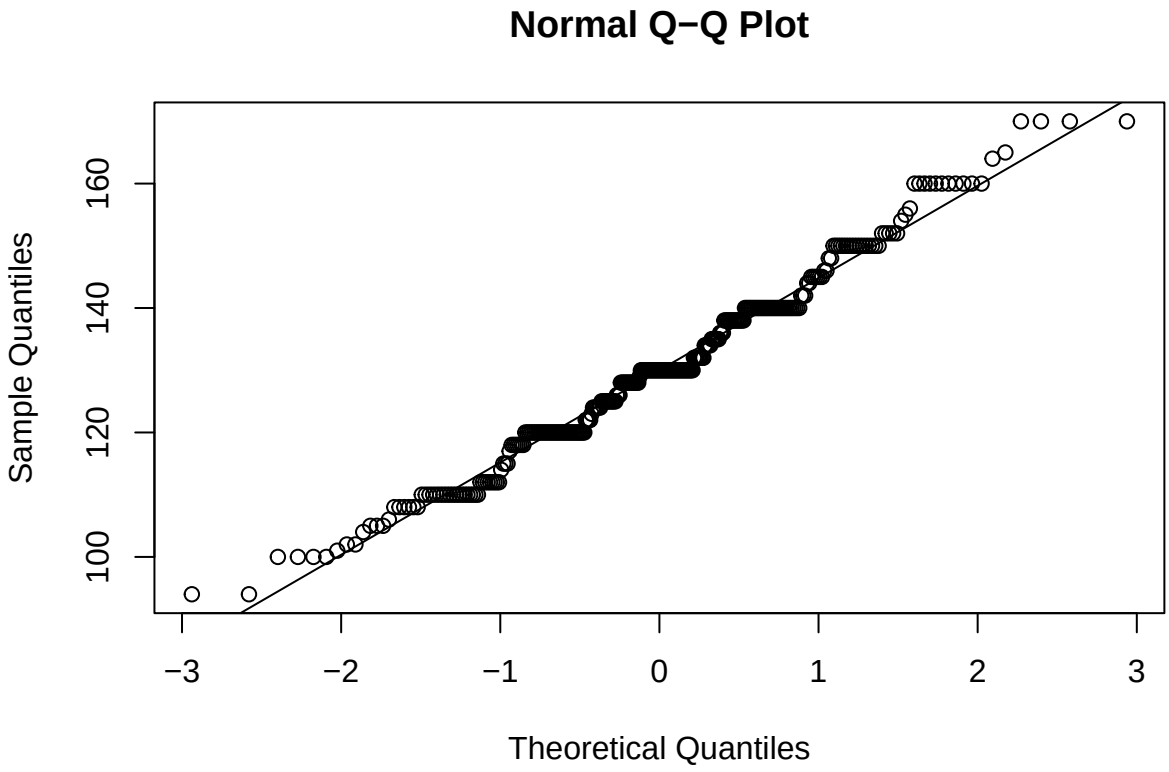
```
qqnorm(heart_sel$age)
qqline(heart_sel$age)
```




```
qqnorm(heart_sel$chol)
qqline(heart_sel$chol)
```

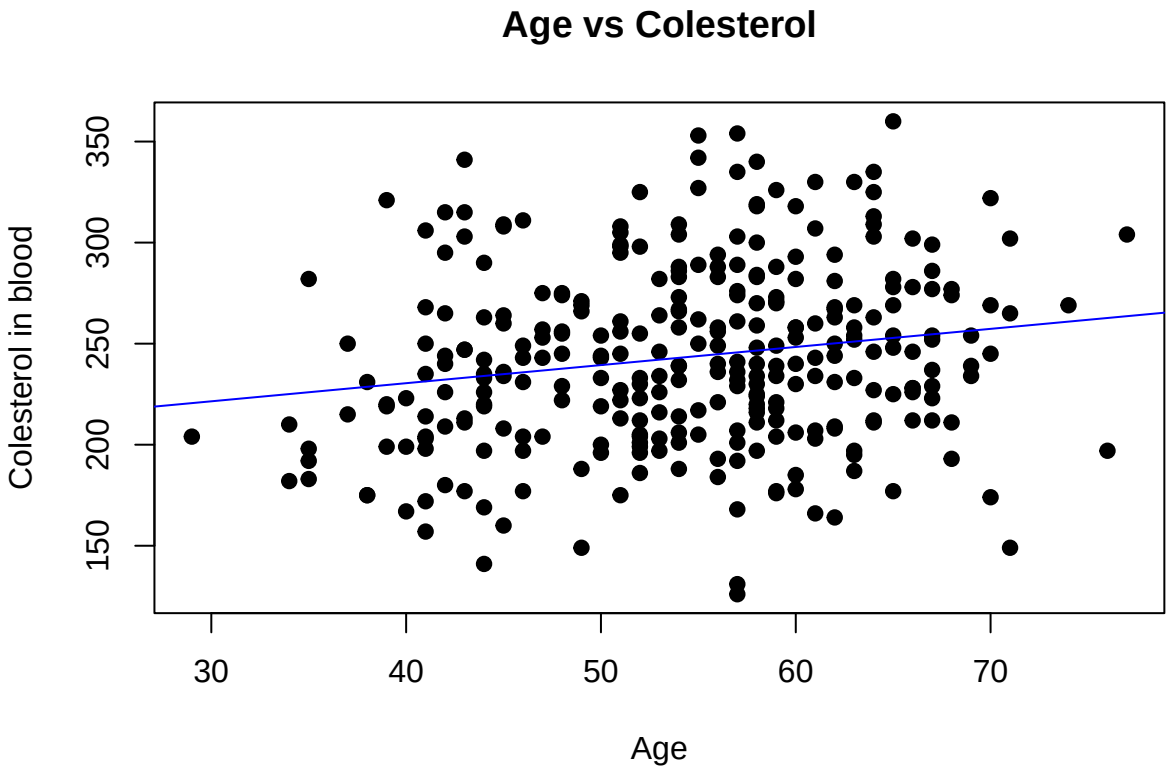



```
qqnorm(heart_sel$trtbps)
qqline(heart_sel$trtbps)
```



Tots els tests ni normogrames QQ ens indiquen que la distribució de les dades no és normal i per tant aplicarem per veure la correlació entre elles el test de rangs de kendall per a variables de distribuió no normal i que a més tingui capacitat de treballar si dos ocurrències cauen dintre el mateix rang.

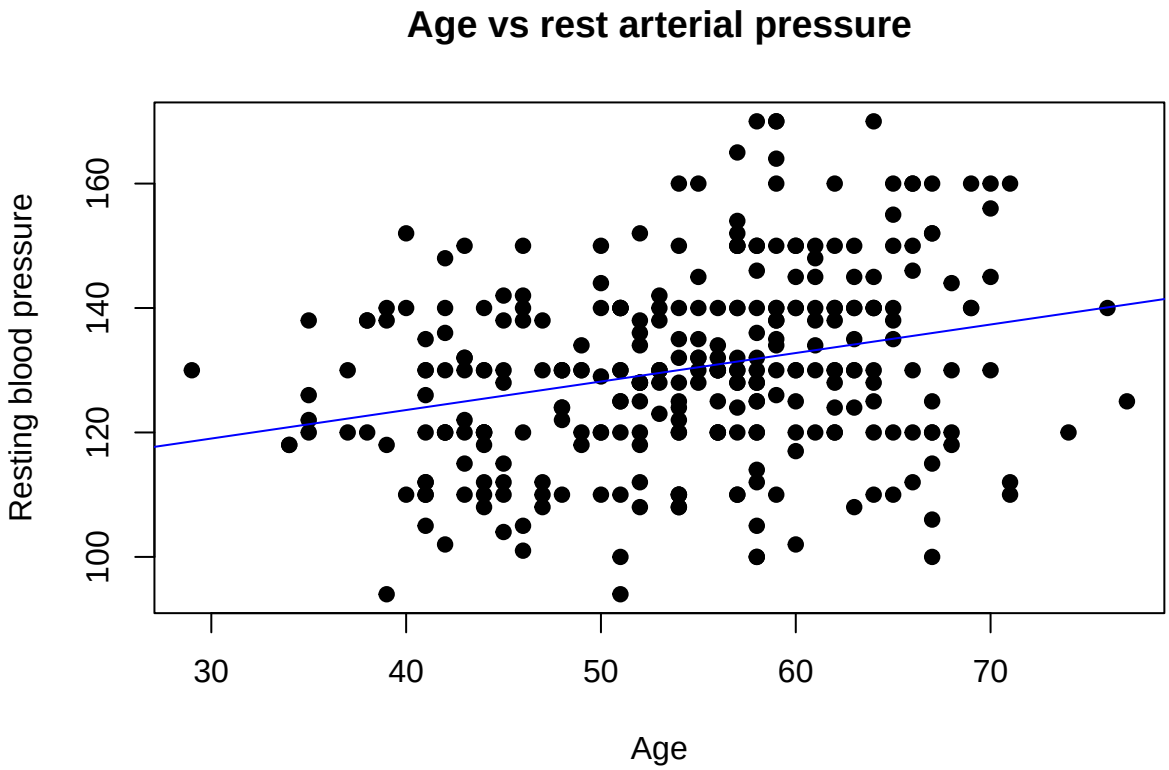
```
# La correlació de rangs de Kendall o Tau s'utilitza per estimar la mesura de l'associació de dues vari

plot(heart_sel$age, heart_sel$chol, main="Age vs Colesterol",
xlab="Age", ylab="Colesterol in blood", pch=19)
abline(lm(chol ~ age, data = heart_sel), col = "blue")
```




```
cor.test(heart_sel$age, heart_sel$chol,method = 'kendall',use = "complete.obs")
```

```
##
##  Kendall's rank correlation tau
##
## data:  heart_sel$age and heart_sel$chol
## z = 3.1499, p-value = 0.001634
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1234615
```

```
plot(heart_sel$age, heart_sel$trtbps, main="Age vs rest arterial pressure",
xlab="Age", ylab="Resting blood pressure", pch=19)
abline(lm(trtbps ~ age, data = heart_sel), col = "blue")
```




```
cor.test(heart_sel$age, heart_sel$trtbps,method = 'kendall',use = "complete.obs")
```

```
##
##  Kendall's rank correlation tau
##
## data:  heart_sel$age and heart_sel$trtbps
## z = 4.8575, p-value = 1.189e-06
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.1952669
```

Resultats i conclusion: Tant en els gràfics com en els tests s'observa que tant la correlació entre edad i colesterol com edad i pressió arterial en repòs no estan associades com evidencia un gràfic amb distribució tipus núvol i una tau de 0.1.

```
#Performs a Fligner-Killeen (median) test of the null that the variances in each of the groups (samples
fligner.test(age ~ sex, data = heart_sel)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  age by sex
## Fligner-Killeen:med chi-squared = 0.517, df = 1, p-value = 0.4721
```

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  trtbps by sex
## Fligner-Killeen:med chi-squared = 0.007459, df = 1, p-value = 0.9312
```

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chol by sex
## Fligner-Killeen:med chi-squared = 1.8065, df = 1, p-value = 0.1789
```

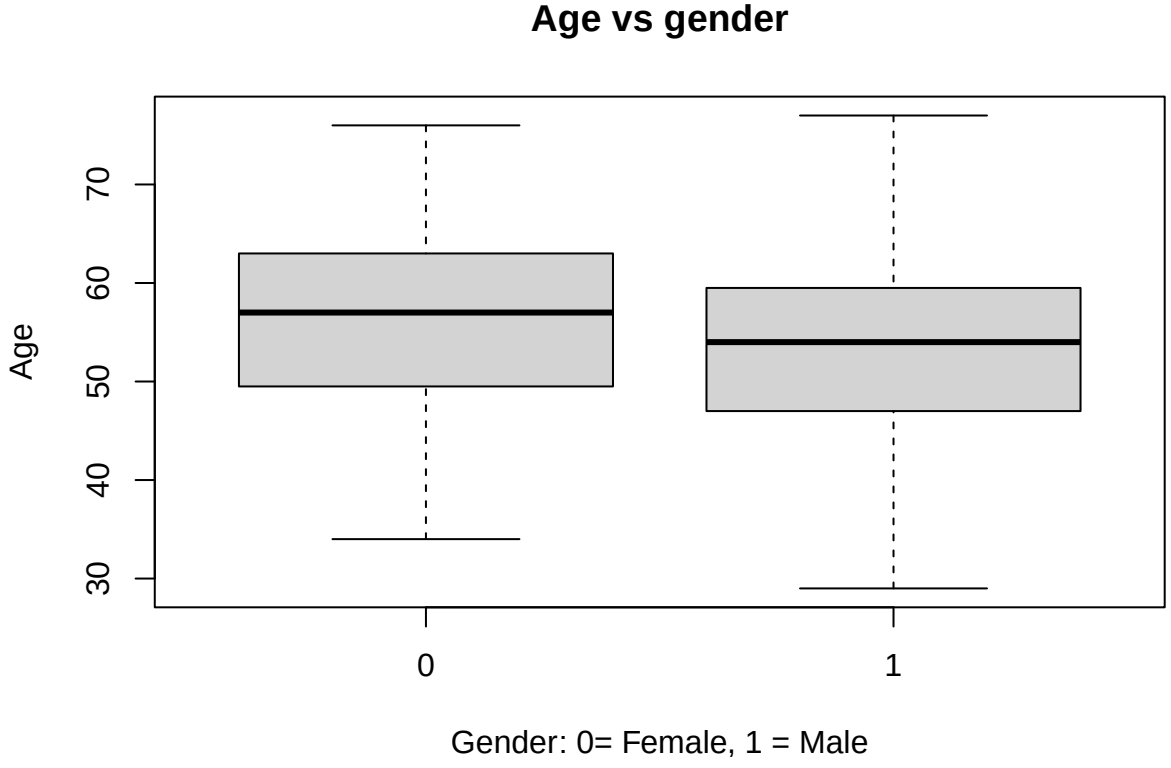
Conclusió: Les distribucions de les variables continues només la presió arterial en repós presenta homocedasticitat.

3.- Comparar si homes i dones tenen un nivell de colesterol i pressió arterial igual o diferent.

Nota important: Cap de les variables continues té criteris per poder aplicar la t-student perquè no partim de distribucions normals i dues d'elles compleixen el criteri d'homocedasticitat.

Per tant utilitzarem el teorema del límit central per poder fer una aproximació paramètrica. El teorema del límit central ens deixa assumir normalitat enla distribució de les mitges mostrals obtingudes de poblacions grans(>30 individus.)

```
# Previament verificarem si l'edat d'homes i dones d'aquesta cohort es comparable i per tant les diferè


# Comparació de les edats pels diferents sexes.

boxplot(heart_sel$age ~ heart_sel$sex, main="Age vs gender",
        xlab="Gender: 0= Female, 1 = Male", ylab="Age")
```




```
mitges<- with(heart_sel,tapply(age,sex,mean))
desvest<-with(heart_sel,tapply(age,sex,sd))
n<-with(heart_sel,tapply(age,sex,length))
```

```r
SE.1 <-  desvest[1] / sqrt(n[1])
SE.2 <-  desvest[2] / sqrt(n[2])
SE      <- sqrt( SE.1^2 + SE.2^2)
z       <- (mitges[1] - mitges[2]) / SE
P.z     <- pnorm(z, lower.tail = FALSE)
print("Desviació estandard")
```

```
## [1] "Desviació estandard"
```

```r
print(SE)
```

```
##        0
## 1.141719
```

```r
print("Diferencia de mitges entre grups")
```

```
## [1] "Diferencia de mitges entre grups"
```

```r
print(mitges[1]-mitges[2])
```

```
##        0
## 1.918629
```

```r
print("estadístic Z")
```

```
## [1] "estadístic Z"
```

```r
print(z)
```

```
##        0
## 1.680474
```

```r
print("probabilitat d'equivocar-nos en rebutjar siguent certa la hipotesis nul·la")
```

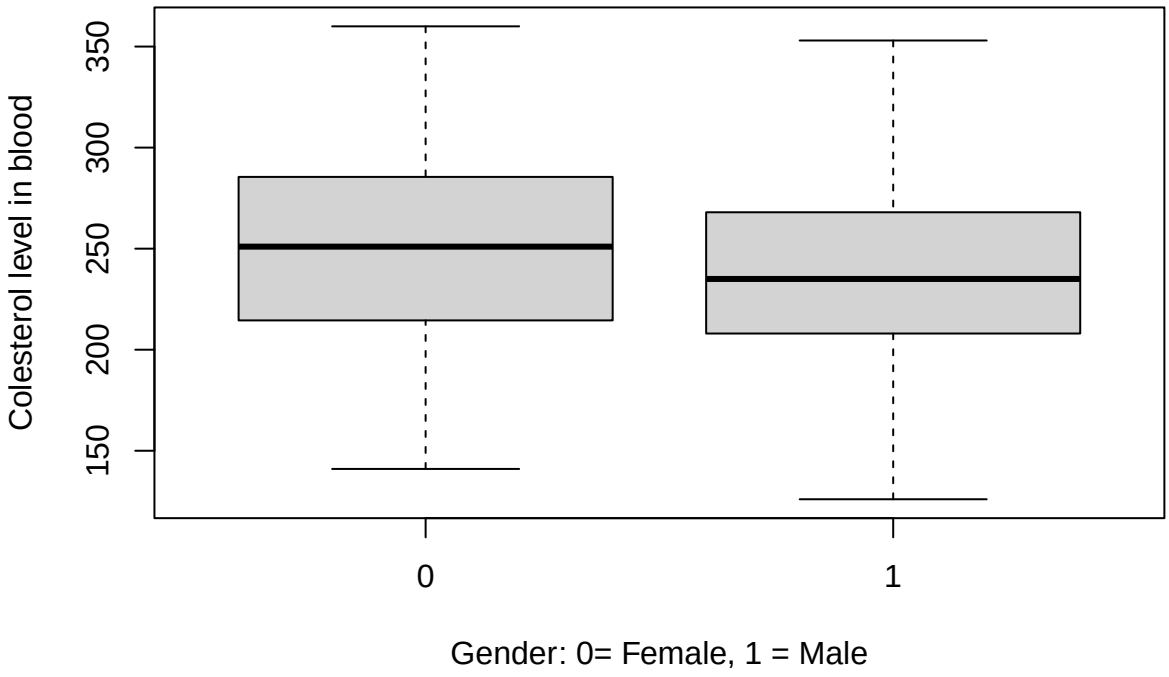
```
## [1] "probabilitat d'equivocar-nos en rebutjar siguent certa la hipotesis nul·la"
```

```r
print(P.z)
```

```
##          0
## 0.04643261
```

Resposta conclusions: El test ens assegura que si rebutgem la hipotesis nul.la i per tant assumim que l'edat d'homes i dones és diferent en la població d'origen, tenim molt poques probabilitats d'equivocar-nos. Per tant en els anàlisi posteriors d'aquesta secció hem de tenir en compte que una variable confusora pot ser l'edat ja que està distribuida de manera diferent entre homes i dones.

```r
# Diferencies en nivells de colesterol entre homes i dones.


boxplot(heart_sel$chol ~ heart_sel$sex, main="Colesterol vs gender",
        xlab="Gender: 0= Female, 1 = Male", ylab="Colesterol level in blood")
```




```
mitges<- with(heart_sel,tapply(chol,sex,mean))
desvest<-with(heart_sel,tapply(chol,sex,sd))
n<-with(heart_sel,tapply(chol,sex,length))

SE.1 <-  desvest[1] / sqrt(n[1])
SE.2 <-  desvest[2] / sqrt(n[2])
SE      <- sqrt( SE.1^2 + SE.2^2)
z       <- (mitges[1] - mitges[2]) / SE
P.z     <- pnorm(z, lower.tail = FALSE)
print("Desviació estandard")
```

```
## [1] "Desviació estandard"
```

```
print(SE)
```

```
##        0
## 5.723592
```

```
print("Diferencia de mitges entre grups")
```

```
## [1] "Diferencia de mitges entre grups"
```

```
print(mitges[1]-mitges[2])
```

```
##        0
## 12.68931
```

```
print("estadístic Z")
```

```
## [1] "estadístic Z"
```

```
print(z)
```

```
##        0
## 2.217019
```

```
print("probabilitat d'equivocar-nos en rebutjar siguent certa la hipotesis nul·la")
```

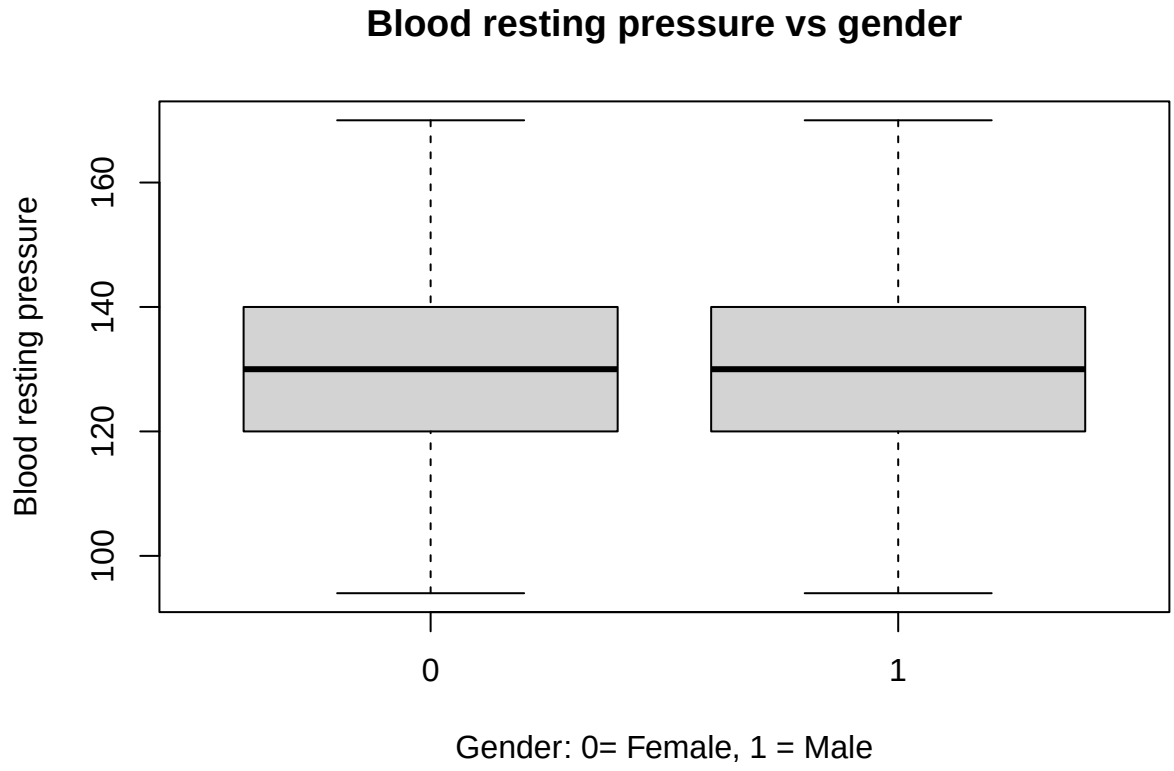
```
## [1] "probabilitat d'equivocar-nos en rebutjar siguent certa la hipotesis nul·la"
```

```
print(P.z)
```

```
##         0
## 0.0133109
```

Resposta i conclusions: El test ens diu que podem rebutjar la igualtat de nivells de colesterol entre homes i dones. Però com hem comentat en l'apartat anterior l'edat podria contribuir al nivell de colesterol més alt en les dones perque en el dataset son més grans.

```
# Diferencies en nivells en presió arterial en repòs entre homes i dones.

boxplot(heart_sel$trtbps ~ heart_sel$sex, main="Blood resting pressure vs gender",
        xlab="Gender: 0= Female, 1 = Male", ylab="Blood resting pressure")
```




```
mitges<- with(heart_sel,tapply(trtbps,sex,mean))
desvest<-with(heart_sel,tapply(trtbps,sex,sd))
n<-with(heart_sel,tapply(trtbps,sex,length))

SE.1 <-  desvest[1] / sqrt(n[1])
SE.2 <-  desvest[2] / sqrt(n[2])
SE      <- sqrt( SE.1^2 + SE.2^2)
z       <- (mitges[1] - mitges[2]) / SE
P.z     <- pnorm(z, lower.tail = FALSE)

print("Desviació estandard")
```

```
## [1] "Desviació estandard"
```

```
print(SE)
```

```
##        0
## 1.888996
```

```
print("Diferencia de mitges entre grups")
```

```
## [1] "Diferencia de mitges entre grups"
```

```
print(mitges[1]-mitges[2])
```

```
##         0
## 0.6242452
```

```
print("estadístic Z")
```

```
## [1] "estadístic Z"
```

```
print(z)
```

```
##         0
## 0.3304641
```

```
print("probabilitat d'equivocar-nos en rebutjar siguent certa la hipotesis nul·la")
```

```
## [1] "probabilitat d'equivocar-nos en rebutjar siguent certa la hipotesis nul·la"
```

```
print(P.z)
```

```
##         0
## 0.3705247
```

Resposta: En aquest test no hi ha diferencies entre els nivells de pressió arterial en repòs.

Regressió Logística

```
# regressió logistica

# canvi d'etiqueta per millorar llegibilitat de la regressió
levels(heart_sel$exng) <-c('no_pain','pain')
levels(heart_sel$sex) <-c('female','male')


mod1 <- glm(output ~ age+sex+trtbps+chol+fbs+restecg+exng+caa, family = binomial, data = heart_sel)
summary(mod1)
```

```
##
## Call:
## glm(formula = output ~ age + sex + trtbps + chol + fbs + restecg +
##     exng + caa, family = binomial, data = heart_sel)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0935  -0.7329   0.2995   0.7014   2.1404
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.032249   1.836601   3.829 0.000129 ***
## age         -0.038257   0.018742  -2.041 0.041226 *
```

```
## sexmale     -1.594058   0.357115  -4.464 8.06e-06 ***
## trtbps      -0.010073   0.010444  -0.964 0.334811
## chol        -0.005760   0.003508  -1.642 0.100631
## fbs1         0.519860   0.416561   1.248 0.212038
## restecg1     0.576661   0.302212   1.908 0.056374 .
## exngpain    -2.101385   0.330181  -6.364 1.96e-10 ***
## caa         -0.844719   0.163518  -5.166 2.39e-07 ***
## ---
## 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: 280.54  on 294  degrees of freedom
## AIC: 298.54
##
## Number of Fisher Scoring iterations: 5
```

```
# Càlcul de les OR (Odds-Ràtio)
exp(coefficients(mod1))
```

```
##  (Intercept)          age      sexmale       trtbps         chol         fbs1
## 1132.5751932    0.9624654    0.2030998    0.9899778    0.9942568    1.6817919
##     restecg1     exngpain          caa
##    1.7800845    0.1222870    0.4296781
```

Hem generat una regresió logística utilitzant com a variables independents l'edat, la presió arterial, els nivells de colesterol i vasos calcificats com a variables continues i glucosa elevada, el sexe, angina o dolor durant esforç i alteracions del ECG com a variables discretes. Com a variable target hem utilitzat outcome que es dicotomica i representa percentatge d'obstrucció de les artèries coronaries més o menys de 50%. A partir del 50% es considera un risc molt elevat de patir un infart de miocardi per obstrucció. El motiu de la regressió és veure quines son les variables més importants a la hora de predir el risc d'infart de miocardi. Un seguit de variables han resultat significatives i per tant amb incidència sobre la predicció del resultat. Son les següents: Edat, Sexe, Dolor anginós en l'esforç i Nombre d'arteries coronaries calcificades.

Certament la interpretació mèdica d'aquests resultats és difícil ja que son contra intuitius:

- EDAT:Al augmentar l'edat sembla que hi ha un mínim factor protector ja el odds (obtingut al exponenciar el log(odds)) li confereix un descens del 4% per any de l'individu. Aquest resultat no s'ajusta a la realitat de la població general. Però pot ser que al ser el dataset de pacients que tenen simptomatologia compatible amb cardiopatia isquèmica, la gent més gran tingui més possibles diagnòstics diferencials al tenir potencialment més malalties.
- SEXE: En aquesta regressió tenir el sexe masculí rebaixa un 80%(odds ratio 0.2) les probabilitats de tenir malaltia coronària respecte al sexe femení. Novament aquest resultat no s'ajusta a la població general on hi ha més incidència d'infarts en homes. Una explicació podria ser que les dones en aquest dataset son més grans però ho descartem ja que a la regressió s'hi inclou la edat per tant aquest possible factor de confusió queda descartat.
- Tant tenir dolor anginós en l'esforç com tenir les artèries calcificades en una fluoroscopia (avui en dia ja no s'utilitza) en general s'associa a més risc de tenir malaltia coronària. En el cas que ens ocupa sembla que per cada artèria calcificada es redueix un 50% el risc de coronariopatia i tenir dolor a l'exercici reduix un 80% les possibilitats de malaltia coronària respecte a tenir no tenir dolor.

La conclusió es que en el context dels resultats creiem que la població del dataset és molt específica i no s'assembla al comportament general. Amb més dades sobre quina és aquesta població concreta podríem afinar més l'explicació dels resultats.

5. Representació dels resultats a partir de taules i gràfiques. Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica.

Les gràfiques estan insertades dintre de cada apartat del document.

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

Els comenetaris sobre resultats i conclusions estan insertades dintre de cada apartat del document.

```
# Creació d'un document csv amb el dataset transformat.


write.csv(heart_sel, "heart_transformed.csv", row.names=FALSE)
```