

## Pràctica 2: Tipologia i cicle de vida de les dades

Estrella Fernàndez i Anna Corral

### 1. Descripció del dataset.

El dataset conté la informació rellevant de diversos pacients en relació a la seva edat i altres paràmetres relacionats amb la seva salut i la probabilitat de tenir un atac al cor. És important aquesta informació per poder correlacionar quines variables del dataset poden estar relacionades amb el fet que el pacient tingui major probabilitat de patir un atac al cor, com per exemple, l'edat, el sexe, els nivells de colesterol en sang, etc.

Per tant, la pregunta o problema que es pretén respondre és quina o quines d'aquestes variables influeix més en que el pacient pateixi d'un atac al cor.

Normalment les malalties relacionades amb el cor es diagnostiquen estudiant els senyals elèctrics del cor per mitjà d'electrocardiogrames per poder determinar com de ràpid batega el cor. També s'estudia a partir de dades de com reacciona el cor quan es sotmès a un estrès o a alguna activitat física.

Les malalties cardiovasculars són la principal causa de mort al nostre planeta amb un total de 17.9 milions de morts cada any. Aquestes, són causades principalment per la hipertensió, la diabetis, el sobrepès i en conclusió, els estils de vida no saludables.

El nostre propòsit és poder prevenir una possible malaltia que afecti al cor. És per això, que l'objectiu del nostre anàlisi és poder construir un model estadístic que sigui capaç d'identificar les variables de dataset presentat a l'enunciat. Per tant, la nostra intenció és poder ser capaces d'identificar factors i la influència que tenen aquests sobre les malalties cardiovasculars que poden arribar a provocar un atac. D'aquesta manera, també serem capaços de facilitar un futur diagnòstic precoç als nostres pacients i potser, poder arribar a preveure morts que siguin produïdes per les mateixes causes al nostre estudi.

### 2. Integració i selecció de les dades d'interès a analitzar.

Com podrem observar en les següents línies de codi, el primer que hem fet després de definir sobre quin dataset volem treballar (el mateix proposat a l'enunciat), és carregar les dades. El primer arxiu .csv l'hem anomenat *heart* i el segon l'hem anomenat *o2*. A l'hora de crear-los, els hem afegit una nova columna anomenada *index* per, posteriorment poder-los unir en un únic fitxer que anomenem *all\_data*. Alhora de carregar el .csv de l'o2 saturation, ens assegurem d'afegir el nom a la columna en la primera fila ja que per defecte ens proposa una dada numèrica.

Per últim, visualitzem les dades del conjunt que acabem de crear per comprovar que s'ha carregat la informació correctament i veiem que es tracta d'un dataset amb 303 files i 16 columnes.

```
> heart <- read.csv("heart.csv")
> heart <- data.frame(heart)
> heart$index <- 1:nrow(heart)

> o2 <- read.csv("o2saturation.csv")
> o2 <- data.frame(o2)
> colnames(o2) <- c("o2_saturation")
> o2$index <- 1:nrow(o2)

> all_data <- merge(heart, o2, by='index', all=TRUE)
> all_data <- data.frame(all_data)
> dim(all_data)
[1] 3585 16
```

```

> head(all_data)
  index age sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa
1     1  63  1  3   145  233   1       0    150    0    2.3   0   0
2     2  37  1  2   130  250   0       1    187    0    3.5   0   0
3     3  41  0  1   130  204   0       0    172    0    1.4   2   0
4     4  56  1  1   120  236   0       1    178    0    0.8   2   0
5     5  57  0  0   120  354   0       1    163    1    0.6   2   0
6     6  57  1  0   140  192   0       1    148    0    0.4   1   0

  thall output o2 saturation
1     1     1      98.6
2     2     1      98.6
3     2     1      98.6
4     2     1      98.1
5     2     1      97.5
6     1     1      97.5

> summary(all_data)
      index      age      sex      cp
Min.   : 1   Min. :29.00   Min. :0.000   Min. :0.000
1st Qu.:897   1st Qu.:47.50   1st Qu.:0.000   1st Qu.:0.000
Median :1793   Median :55.00   Median :1.000   Median :1.000
Mean   :1793   Mean   :54.37   Mean   :0.683   Mean   :0.967
3rd Qu.:2689   3rd Qu.:61.00   3rd Qu.:1.000   3rd Qu.:2.000
Max.   :3585   Max.   :77.00   Max.   :1.000   Max.   :3.000
NA's   :3282   NA's   :3282   NA's   :3282   NA's   :3282

      trtbps      chol      fbs      restecg
Min.   : 94.0   Min. :126.0   Min. :0.000   Min. :0.000
1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.000   1st Qu.:0.000
Median :130.0   Median :240.0   Median :0.000   Median :1.000
Mean   :131.6   Mean   :246.3   Mean   :0.149   Mean   :0.528
3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.000   3rd Qu.:1.000
Max.   :200.0   Max.   :564.0   Max.   :1.000   Max.   :2.000
NA's   :3282   NA's   :3282   NA's   :3282   NA's   :3282

      thalachh      exng      oldpeak      slp
Min.   : 71.0   Min. :0.000   Min. :0.00   Min. :0.000
1st Qu.:133.5   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:1.000
Median :153.0   Median :0.000   Median :0.80   Median :1.000
Mean   :149.6   Mean   :0.327   Mean   :1.04   Mean   :1.399
3rd Qu.:166.0   3rd Qu.:1.000   3rd Qu.:1.60   3rd Qu.:2.000
Max.   :202.0   Max.   :1.000   Max.   :6.20   Max.   :2.000
NA's   :3282   NA's   :3282   NA's   :3282   NA's   :3282

      caa      thall      output      O2 Saturation
Min.   :0.000   Min. :0.000   Min. :0.000   Min. :96.50
1st Qu.:0.000   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:97.60
Median :0.000   Median :2.000   Median :1.000   Median :98.60
Mean   :0.729   Mean   :2.314   Mean   :0.545   Mean   :98.24
3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:98.60
Max.   :4.000   Max.   :3.000   Max.   :1.000   Max.   :99.60
NA's   :3282   NA's   :3282   NA's   :3282

> str(all_data)
'data.frame': 3585 obs. of 16 variables:
 $ index      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ 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 1 1 2 2 2 ...
 $ caa        : int  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 ...
 $ O2_Saturation: num  98.6 98.6 98.6 98.1 97.5 97.5 97.5 97.5 97.5 97.5 ...

```

Agafarem les dades dels dos arxius .csv que trobem al dataset *Heart Attack Analysis & Prediction dataset*. (<https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>).

Per l'arxiu de dades "heart.csv" agafarem les variables següents:

- Age : Age of the patient in years
- Sex : Sex of the patient (0= female; 1= male)

- exang: exercise induced angina (1 = yes; 0 = no)
- ca: number of major vessels (0-3)
- cp : Chest Pain type
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
- trtbps : resting blood pressure (in mm/Hg)
- chol : serum cholesterol in mg/dl fetched via BMI sensor
- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest\_ecg : resting electrocardiographic results
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach : maximum heart rate achieved
- oldpeak: ST depression induced by exercise relative to rest.
- thall: A blood disorder called thalassemia.
  - Value 0: NULL
  - Value 1: fixed defect (no blood flow in some part of the heart)
  - Value 2: normal blood flow
  - Value 3: reversible defect (a blood flow is observed but it is not normal)
- slp: the slope of the peak exercise ST segment
  - Value 0: downsloping
  - Value 1: flat
  - Value 2: upsloping
- target : 0= less chance of heart attack 1= more chance of heart attack

Per l'arxiu de dades "o2Saturation.csv" agafarem la variable següent:

- O2 Saturation: Measurement of arterial oxygen saturation expressed in %.

Un cop tenim els dos arxius .csv d'una manera conjunta i les dades ben definides, optem per fer una representació gràfica de cadascuna de les variables mitjançant una funció anomenada *data* que ens recorre totes les columnes excepte la d'*index* i ens que ens mostra la distribució de cada variable. Amb aquesta representació també ens podrem arribar a fer una idea de si tenim en el nostre dataset algun valor extrem del que haguem de prescindir més endavant ja que ens ajuda a veure-ho d'una manera més visual i clara.

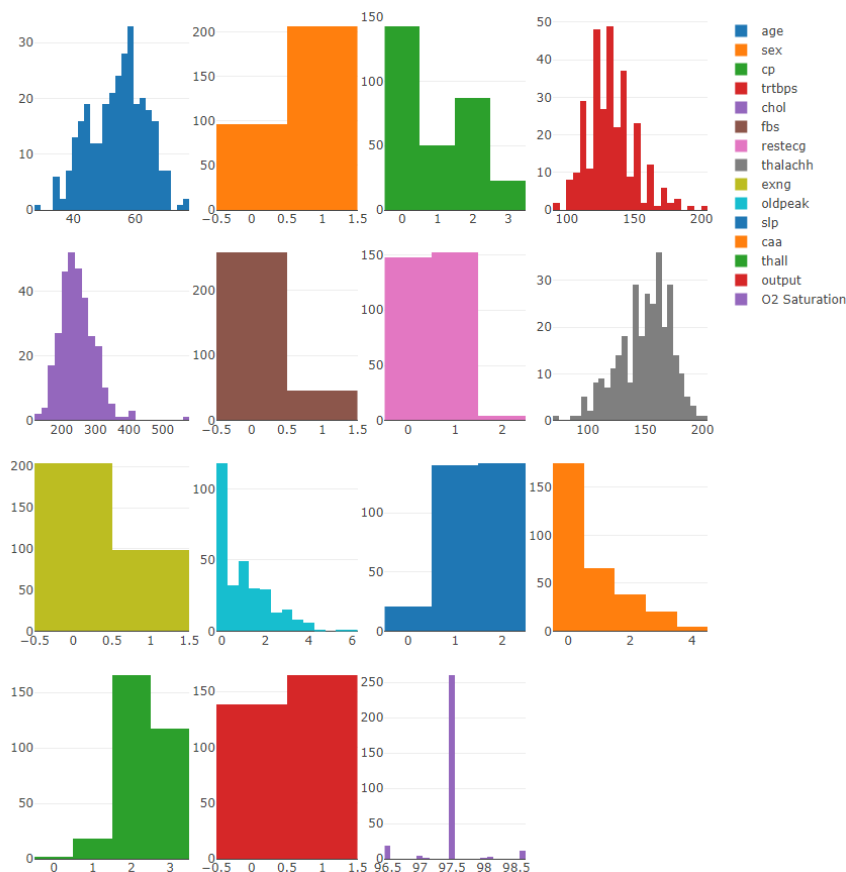
Primerament, després de definir la funció, creem una llista buida dels gràfics per, a continuació, seleccionar quines són les columnes que recorrem, en concret de la 2 a la 16 obviat l'*index*. Després, creem un *for* que ens passi per totes les variables i guardem la representació en la llista buida ja creada anteriorment. Per últim, especifiquem que volem la representació en 4 files i cridem a la funció sobre el dataset que ens interessa representar, en aquest cas sobre el *all\_data*.

La funció descrita té el següent aspecte:

```

> plot_data <- function(data){
+   # Creació d'una llista de gràfics
+   plots <- list()
+   names <- colnames(data[2:16])
+   # Passem per totes les variables i guardem la representació
+   # en la llista creada anteriorment
+   for (i in 1:length(names)){
+     fig <- plot_ly(x = data[,i+1],
+                   type = "histogram",
+                   histnorm = "count",
+                   name = names[i])
+     plots[[i]] <- fig
+   }
+   # Representem els gràfics en 4 files
+   representacio <- subplot(plots, nrows = 4)
+   representacio
+ }
> plot_data(all_data)

```



(La visualització d'aquestes dades en Rstudio és interactiva podent posar el cursor sobre qualsevol espai dels gràfics i que ens mostri la dada sobre la que ens trobem específicament) D'aquesta representació podem treure com a primera conclusió molt general que no trobem valors extrems ni valors buits que ens dificultin potser l'enteniment correcte de les dades. Tot i així, ho estudiem a fons en el següent apartat.

### 3. Neteja de les dades.

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

Per poder respondre a aquesta pregunta, ens fixem en quin tipus de variable és cadascuna de les categories que ens conformen el dataset.

Nom abreviat	Descripció	Valors
age	Edat de la persona en anys	
sex	Gènere	0: female, 1: male
cp	Chest Pain Type	0 = asymptomatic, 1 = atypical angina, 2 = non-anginal pain, 3 = typical angina
trtbps	Person's resting blood pressure (in mm Hg)	valor numèric
chol	Person's cholesterol (in mg/dl)	valor numèric
fbs	Person's fasting blood sugar (if > 120 mg/dl)	0 = false, 1 = true
restecg	Resting electrocardiographic result	0 = probable or definite left ventricular hypertrophy, 1 = normal, 2 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
thalachh	Person's maximum heart rate achieved	valor numèric
exng	Exercise induced angina	0 = no, 1 = yes
oldpeak	ST depression induced by exercise relative to rest	valor numèric
slp	Slope of the peak exercise ST segment	0 = downsloping, 1 = flat, 2 = upsloping
caa	Number of major vessels	Valors entre el 0-4
thall	A blood disorder called thalassemia	0 = NA's, 1 = fixed defect, 2 = normal, 3 = reversable defect
output	Target, Heart disease	0 = no, 1 = yes
o2Saturation	Saturation level	valor numèric

Com podem observar, algunes de les variables, com per exemple; *cp*, *fbs*, *restecg*, *exng*, *slp*, *caa* i *output*, contenen zeros perquè es tracta de variables categòriques on el 0 està assignat a algun tipus de subcategoria que defineix la variable en qüestió. En altres casos, com per exemple en

la variable *thall*, ens trobem amb zeros que representen valors buits i que, per tant, eliminem perquè no ens aporten cap informació necessària.

Després de fer els canvis necessaris que acabem de definir, comprovem el nostre dataset fent:

```
> all_data<- na.omit(all_data)
> dim(all_data)
[1] 303 16
> colSums(all_data=="")
      index      age      sex      cp      trtbps
      0      0      0      0      0
      chol      fbs      restecg      thalachh      exng
      0      0      0      0      0
      oldpeak      slp      caa      thall      output
      0      0      0      0      0
O2 Saturation
      0
> colSums(is.na(all_data))
      index      age      sex      cp      trtbps
      0      0      0      0      0
      chol      fbs      restecg      thalachh      exng
      0      0      0      0      0
      oldpeak      slp      caa      thall      output
      0      0      0      0      0
O2 Saturation
      0
```

### 3.2 Identifica i gestiona els valors extrems.

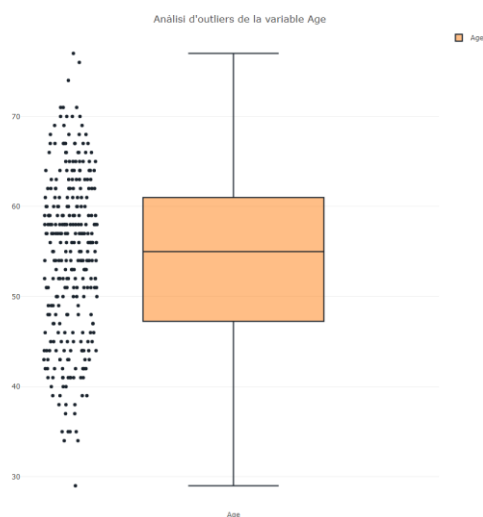
Ahora d'identificar i gestionar els valors extrems que hi puguem trobar, ho hem dividit segons el tipus de dades que estiguem estudiant, segons sigui numèrica o categòrica. Per aquelles que són de caràcter numèric, ens centrarem amb una classificació gràfica en forma de boxplot i per aquelles categòriques farem una representació en forma d'histograma. Hem escollit aquestes representacions perquè creiem que són les més adients per poder extreure la màxima informació útil.

Pel que fa a les **variables numèriques**, hem definit una funció anomenada *anàlisi\_outliers* on seguidament podrem especificar quina variable cridar per poder-la representar. Un cop hem definit aquesta funció, passem a la creació del gràfic on especifiquem de quin tipus és el que volem utilitzar. Seguidament, marquem els paràmetres gràfics d'aquesta representació així com el títol de la figura en qüestió. Finalment, obtenim els valors dels possibles outliers que puguem observar en la nostra figura.

```
> # Boxplot per a les variables numèriques
> analisis_outliers <- function(variable, name){
+
+   # Creació del gràfic
+   fig <- plot_ly(type = 'box')
+
+   # Representació de la variable
+   fig <- fig %>% add_boxplot(y = variable,
+                               jitter = 0.3,
+                               pointpos = -1.8,
+                               boxpoints = 'all',
+                               marker = list(color = 'rgb(23,32,42)'),
+                               line = list(color = 'rgb(23,32,42)'),
+                               name = name)
+
+   fig <- fig %>% layout(title = paste("Anàlisi d'outliers de la variable", name))
+
+   # Obtenció dels possibles outliers
+   outliers <- boxplot.stats(variable)$out
+
+   return(list(outliers=outliers, fig=fig))
+ }
```

Comencem amb la variable *age*:

```
> analisis = analisis_outliers(all_data$age,"Age")
> analisis$fig
```



El que podem observar és que l'edat dels pacients del nostre dataset està compresa entre els 29 i els 77 anys agafant diferents franges d'edat perquè els resultats siguin el més competents possible. L'edat mitjana es troba als 55 anys, el primer quartil als 47 i el tercer quartil als 61. Veiem que la major part de distribució dels punts es centra en la franja dels 50 als 63 anys aproximadament i no trobem cap valor que se surti de la norma.

Passem a la variable *trtbps*:

```
> analisis = analisis_outliers(all_data$trtbps,"Resting blood pressure")
> analisis$fig
> analisis$outliers
[1] 172 178 180 180 200 174 192 178 180
```

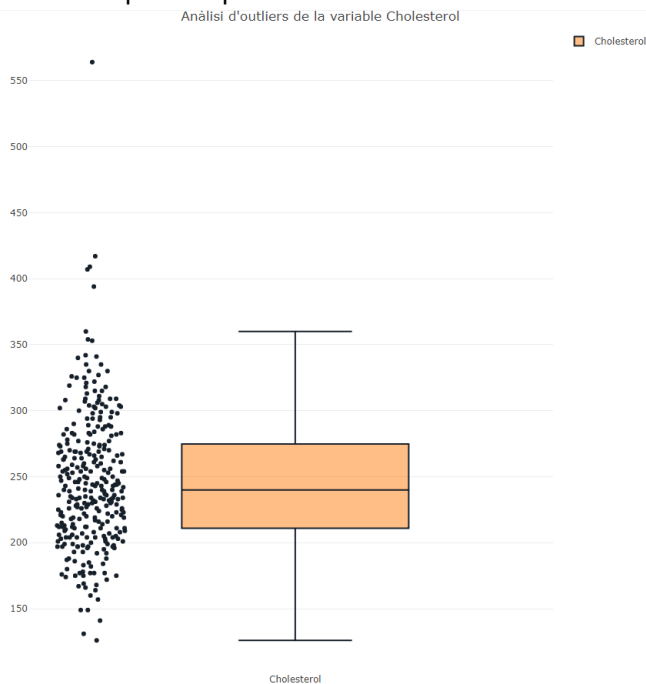
Pel que fa a aquesta variable, veiem que es troba entre un valor mínim de 94 i un valor màxim de 200 on la mitjana esta als 130. El primer quartil el situem als 120 i el tercer als 140. Veiem també que la major concentració de punts coincideix amb els quartils i si que podem observar que hi ha alguns valors que podrien sortir-se de la norma. Mitjançant la tercera línia de codi, veiem com ens extreu aquests valors del que podríem sospitar. Fent recerca sobre la *Resting blood pressure*, veiem que aquests valors poden ser reals i que ens estan indicant que ens troben enmig d'una *Hipertensiva crisis*. D'aquesta manera, tot i que en un principi els haguem pogut interpretar com a outliers, els considerem part de l'estudi.



Passem a la variable *Cholesterol*:

```
> analisis = analisis_outliers(all_data$chol,"Cholesterol")
> analisis$fig
> analisis$outliers
[1] 417 564 394 407 409
```

El que podem observar en aquesta variable és que el valor mínim el trobem al 126 i el màxim als 564 sent la mitjana els 240. El primer quartil es troba als 211 i el tercer als 274 sent entre aquests dos punts on es troba la major distribució de punts del gràfic. El que podem veure gràficament és com molt clarament 5 punts es surten de la distribució “normal”, cosa que podríem considerar un outlier. Tot i així, investiguem mèdicament que poden significar aquestes dades i arribem a entendre que són valors molt extrems i perillosos però que són valors reals que mostren una hipercolesterolèmia molt greu que pot derivar en malalties com la pancreatitis aguda. Per tant, decidim no eliminar ni imputar a la mitja aquesta valors ja que ens donen una informació també valuosa i important pel nostre estudi.



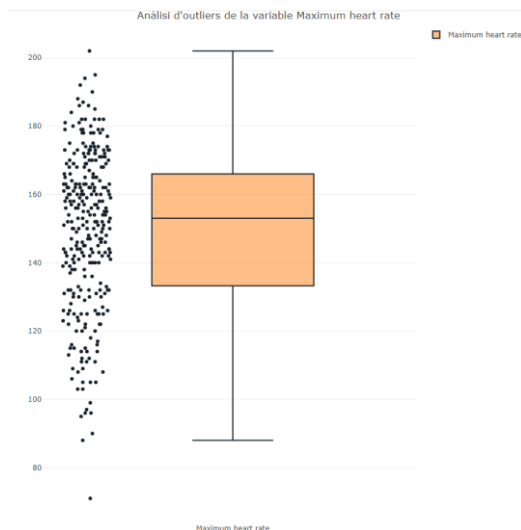
Passem a la variable *thalachh*:

```
> analisis = analisis_outliers(all_data$thalachh,"Maximum heart rate")
> analisis$fig
```

El que podem observar de l'anàlisi de la variable *thalachh* és que el valor mínim és 71 mentre que el màxim és de 202. La mitjana la trobem al valor 153, el primer quartil es troba als 133.25 i el tercer quartil als 166.

Com podem observar a la figura, no s'aprecien valors que surtin de la norma i és per això que no invoquem els possibles valors outliers per interpretar els resultats. Tot i així, investigant els resultats obtinguts, podem establir que el mínim trobat (71) correspon a una freqüència cardíaca en repòs normal per un adult, ja que es troba entre les 60 i les 100 pulsacions per minut que és el llinar del repòs. En canvi, si ens fixem amb el màxim trobat, les 202 pulsacions per minut, corresponen a un estat d'arrítmia generat per la rapidesa dels batecs que poden provocar un episodi de taquicàrdia supraventricular (molt comú en persones joves i sanes).

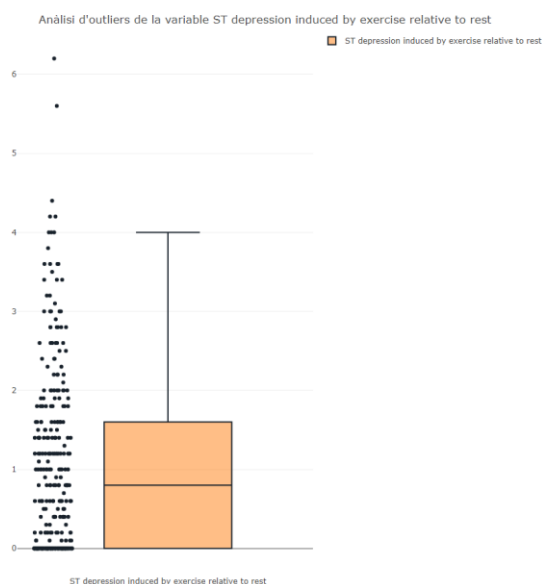




Passem ara a la variable *oldpeak*:

```
> analisis = analisis_outliers(all_data$oldpeak,"ST depression induced by exercise relative to rest")
> analisis$fig
> analisis$outliers
[1] 4.2 6.2 5.6 4.2 4.4
```

El que podem observar d'aquesta representació és que trobem el valor mínim a 0 o el valor màxim a 6.2. La mitjana es troba als 0.8 mentre que el primer quartil és al 0 i el tercer quartil és al 1.6. El que ens està indicant aquesta variable és la probabilitat de patir lesions coronàries que pot provocar inestabilitats cardíques molt greus. D'aquesta manera, observant el gràfic obtingut podem concloure dient que la majoria dels pacients examinats en el nostre dataset es troben en la zona d'una nul·la o molt baixa probabilitat de patir aquesta lesió. Podem veure també com ens surten alguns punts fora de la normal però si investiguem el significat veurem que aquesta variable va del 0 als 6.2, sent aquest últim una certesa absoluta alhora de patir aquestes lesions. Per tant, els 5 punts que ens surten com a outliers es poden sortir de la norma habitual però són dades totalment possibles i importants a tenir en compte en el nostre estudi ja que ens mostra la quantitat de pacients amb una alta probabilitat de patir lesions coronàries.

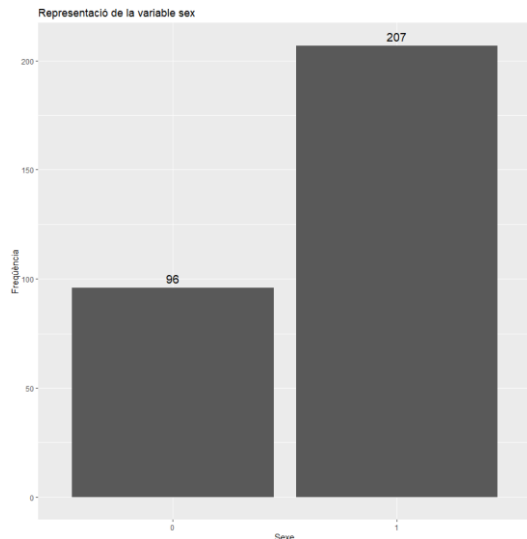


Un cop tenim la representació de les variables numèriques, passem a la representació i interpretació de les variables categòriques.

Comencem per la variable *sex*:

```
> ggplot(all_data,
+   aes(x=factor(sex)))+
+   geom_bar()+
+   labs(title="Representació de la variable sex",
+     x="Sexe", y="Freqüència")+
+   geom_text(aes(label=..count..),stat = 'count',
+     position = position_dodge(0.9),
+     vjust=-0.5,
+     size=5.0)
```

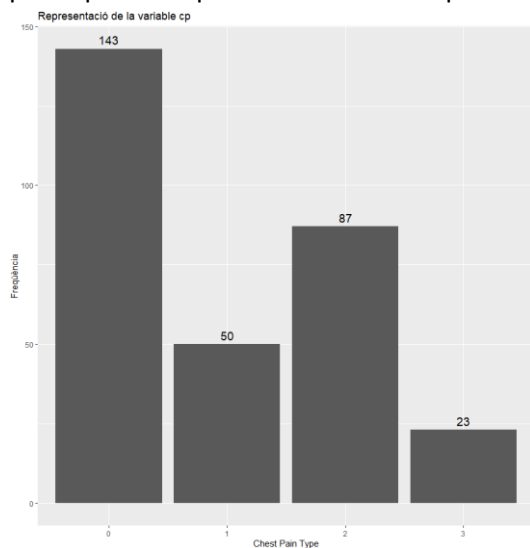
El que podem observar de la representació resultant és quina freqüència d'aparició en el nostre dataset tenim del gènere masculí (1) i de femení (0). Veiem que de 303 observacions, 207 corresponen a homes mentre que 96 corresponen a dones.



Continuem amb la variable *cp*:

```
> ggplot(all_data,
+   aes(x=factor(cp)))+
+   geom_bar()+
+   labs(title="Representació de la variable cp",
+     x="Chest Pain Type", y="Freqüència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+     position = position_dodge(0.9),
+     vjust=-0.5,
+     size=5.0)
```

Veiem que la variable *cp* es divideix en 4 subcategories sent 0: asymptomatic, 1 = atypical angina, 2 = non-anginal pain i 3 = typical angina. El que podem extreure de la representació gràfica és que el cas més freqüent amb 143 pacients és l'asimptomàtic, seguit dels 87 pacients de la categoria non-anginal pain, seguit dels 50 pacients de la categoria atypical angina i, per últim, els de la categoria typical angina amb 23 pacients. Per tant, la major part dels pacients que participen en aquest estudi són asimptomàtics en quant a dolor al pit.

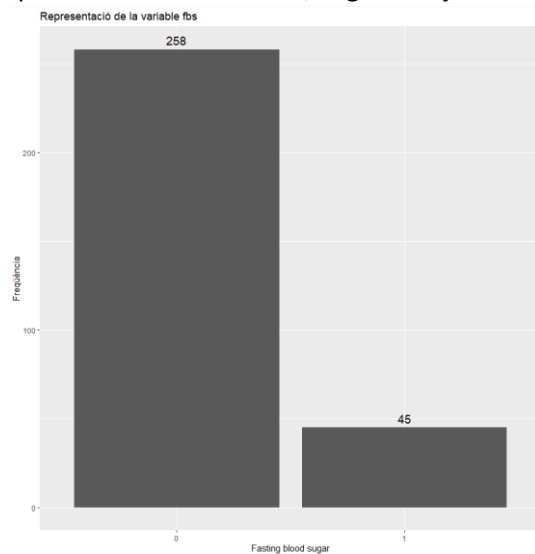


Continuem amb la variable *fbs*:

```
> ggplot(all_data,
+   aes(x=factor(fbs)))+
+   geom_bar()+
+   labs(title="Representació de la variable fbs",
+   x="Fasting blood sugar", y="Freqüència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+   position = position_dodge(0.9),
+   vjust=-0.5,
+   size=5.0)
```

La variable *fbs* està definida per dues subcategories, 0: false i 1:true. Aquesta variable, ens indica que si el *fasting blood sugar* és < a 120, el pacient no pateix diabetis però si aquest valor és > a 120, el pacient en qüestió pateix diabetis.

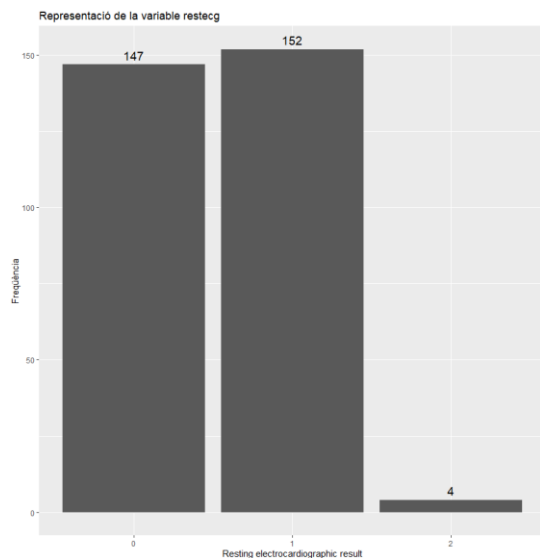
En el nostre cas, tenim 258 pacients que no tenen un *fbs*>120 i que per tant, no pateixen diabetis i després tenim 45 pacients que si tenen un *fbs*>120 i que per tant, tenen diabetis. Podem dir que de 303 observacions, la gran majoria d'elles no pateixen diabetis.



Seguim amb la variable *restecg*:

```
> ggplot(all_data,
+   aes(x=factor(restecg)))+
+   geom_bar()+
+   labs(title="Representació de la variable restecg",
+   x="Resting electrocardiographic result", y="Freqüència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+   position = position_dodge(0.9),
+   vjust=-0.5,
+   size=5.0)
```

Veiem que la variable *restecg* està definida per 3 categories, sent 0: hipertròfia ventricular esquerra, 1 = normal, 2 = anormalitat de l'ona ST-T. El que podem extreure d'aquesta representació és que els grups 0 i 1 són els més comuns en el nostre dataset tenint 147 pacients i 152 pacients que presenten un electrocardiograma normal. En canvi, només tenim 4 casos de la categoria 2 on s'aprecia a l'electrocardiograma una anormalitat en el flux de l'ona.

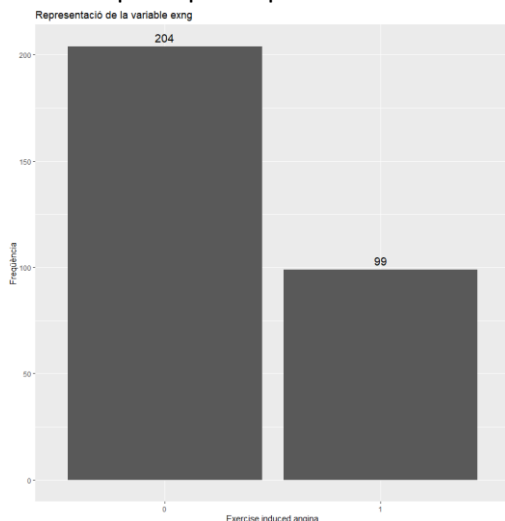


Continuem amb la variable *exng*:

```
> ggplot(all_data,
+   aes(x=factor(exng)))+
+   geom_bar()+
+   labs(title="Representació de la variable exng",
+     x="Exercise induced angina", y="Frequència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+     position = position_dodge(0.9),
+     vjust=-0.5,
+     size=5.0)
```

Ens trobem que aquesta variable està definida per dues categories sent 0: no i 1: yes. L'angina de pit induïda per l'exercici és una dolència comuna per persones amb algun tipus de malaltia cardíaca quan fan exercici en fred ja que sembla ser que aquest fred té efectes adversos i que facilita l'aparició més ràpida de l'angina.

El que ens trobem amb la nostra representació és que 204 pacients no tenen aquest tipus de dolència però que 99 pacients del nostre dataset d'estudi si que la pateixen.



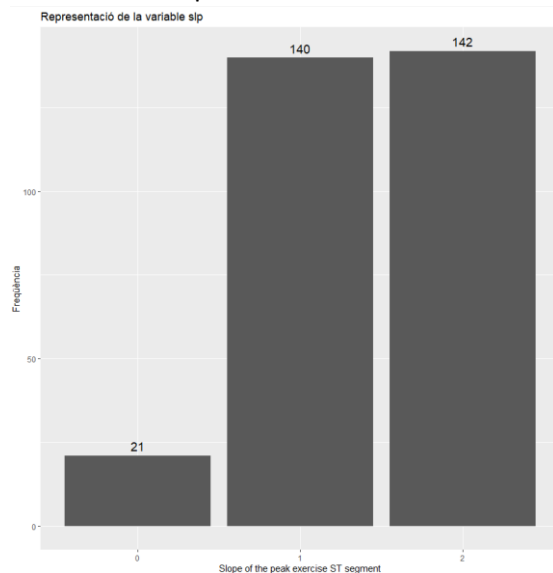
Seguim amb la variable *slp*:

```
> ggplot(all_data,
+   aes(x=factor(slp)))+
+   geom_bar()+
+   labs(title="Representació de la variable slp",
+     x="slope of the peak exercise ST segment", y="Frequència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+     position = position_dodge(0.9),
+     vjust=-0.5,
+     size=5.0)
```

La variable *slp* està definida per 3 categories diferents sent 0: downsloping, 1: flat, 2: upsloping. Aquesta variable el que defineix és l'interval entre la despolarització i la repolarització del

ventricles. Que aquest segment sigui de la categoria 0, downsloping, ens suggereix que hi ha una isquèmia més extensa que la resta. Quan el segment sigui 2, upsloping, ens indica que hi ha isquèmia cardíaca en presència de símptomes cardíacs actius i que el segment sigui 1, flat, ens indica que ens trobem amb un patró de procés moderat.

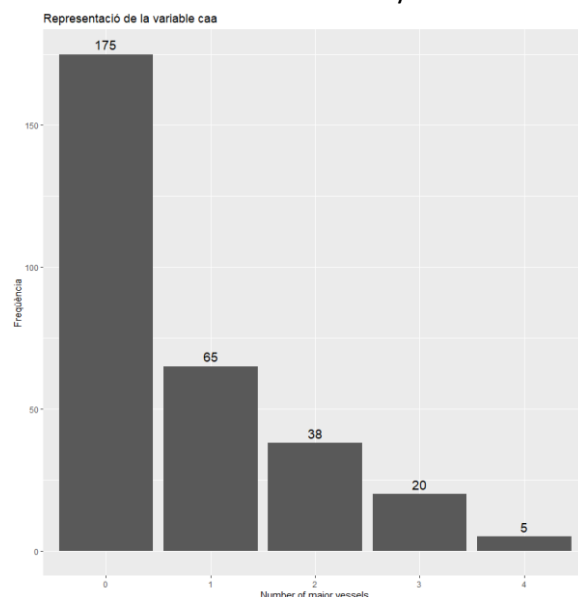
En el nostre cas, ens trobem que tant la categoria 1 com la 2 són les més abundants amb 140 i 142 casos respectivament mentre que la categoria 0 és la menys representada en el nostre grup amb només 21 pacients.



Continuem amb la variable *caa*:

```
> ggplot(all_data,
+       aes(x=factor(caa)))+
+   geom_bar()+
+   labs(title="Representació de la variable caa",
+        x="Number of major vessels", y="Frequència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+             position = position_dodge(0.9),
+             vjust=-0.5,
+             size=5.0)
```

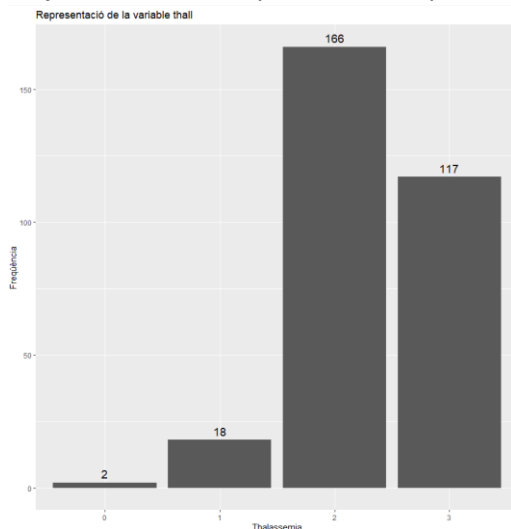
Aquesta variable està definida per 5 categories diferents que poden prendre entre els valors d'entre el 0 i el 4 depenent del nombre de vessels. El que podem observar és que l'abundància de cada categoria segueix un ordre descendent del 0 al 4 sent la categoria 0 la més abundant amb 175 casos i sent la 4 la menys abundant amb només 5 casos.



Continuem amb la variable *thall*:

```
> ggplot(all_data,
+   aes(x=factor(thall)))+
+   geom_bar()+
+   labs(title="Representació de la variable thall",
+     x="Thalassemia", y="Freqüència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+     position = position_dodge(0.9),
+     vjust=-0.5,
+     size=5.0)
```

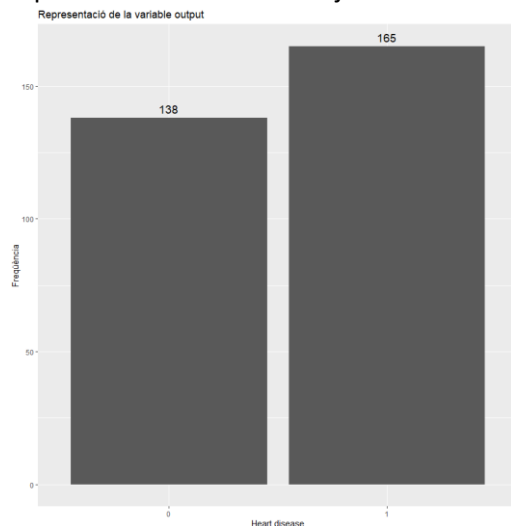
Aquesta variable veiem que està definida per 4 categories diferents que van les 0 al 4 sent 0: null, 1: hi trobem el defecte perquè no hi ha flux de sang en una part del cor, 2: flux normal de la sang i 3: defecte reversible, on s'observa defecte però el flux és normal. Observem que les categories 2 i 3 són les més abundants amb 166 casos i 117 respectivament i que per tant la gran majoria dels nostres pacients no la pateixen o si que la poden presentar però no els afecta.



Finalitzem amb la variable *output*:

```
> ggplot(all_data,
+   aes(x=factor(output)))+
+   geom_bar()+
+   labs(title="Representació de la variable output",
+     x="Heart disease", y="Freqüència")+
+   geom_text(aes(label=after_stat(count)),stat = 'count',
+     position = position_dodge(0.9),
+     vjust=-0.5,
+     size=5.0)
```

Per últim, ens trobem amb la variable objecte del nostre estudi que ens determina si el pacient pot patir un problema cardiovascular amb el número 1 o per la contra, si no el pateix amb el 0. Ens trobem que 138 dels 303 no el pateixen mentre que 165 si que ho fan. Per tant, és una representació molt balancejada entre una variable i l'altre.



#### 4. Anàlisi de les dades i representació dels resultats (apartat 5)

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

Volem analitzar si el gènere de la persona i si la diabetis influeixen sobre les malalties relacionades amb el cor. Per aquest motiu es realitzarà un contrast d'hipòtesi per determinar si:

1. El gènere influeix sobre les malalties del cor
2. La diabetis (nivells de sucre en sang de la persona en dejuni > 120 mg/dl) influeix sobre les malalties del cor

També es realitzarà una matriu de correlacions sobre totes les variables per comprovar si existeix col·linealitat entre les variables explicatives, i així per eliminar-les de l'anàlisi en cas que hi hagués una forta correlació entre les alguna d'aquestes variables.

A més, es realitzarà un model de regressió lineal per analitzar totes les variables dels dos fitxers de dades que contenen informació rellevant al pacient per determinar com aquestes variables poden afectar sobre les malalties del cor, és a dir, com influeixen totes aquestes variables a l'hora de determinar si la persona pateix o no d'alguna malaltia al cor i quines són les que tenen major efecte sobre la variable resposta.

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

Suposarem normalitat en les dades perquè quan la mostra és suficientment gran, la distribució de les dades tendeix a una normal.

#### Anàlisi de la variable OUTPUT en funció de SEX

Podem observar que fent ús de la funció `table()` obtindrem una matriu amb les freqüències. Amb la funció `prop.table()` observem el mateix però amb les proporcions. La funció `CrossTable()` ens mostra tant les freqüències com les proporcions.

```
> tab1a1 <- table(all_data$sex, all_data$output, dnn = c("Sex", "Output"))
> tab1a1
      Output
Sex      0      1
0      24      72
1     114      93
> prop.table(table(all_data$sex, all_data$output))
      0      1
0 0.07920792 0.23762376
1 0.37623762 0.30693069
> CrossTable(all_data$sex, all_data$output, prop.chisq = FALSE)
```

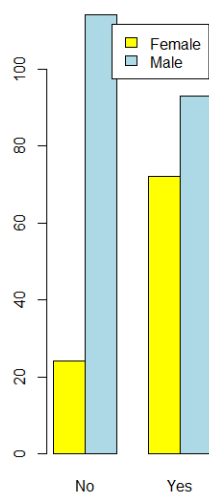
Cell Contents	
	N
N / Row Total	
N / Col Total	
N / Table Total	

Total Observations in Table: 303

all_data\$sex	all_data\$output		Row Total
	0	1	
0	24	72	96
	0.250	0.750	
	0.174	0.436	
	0.079	0.238	
1	114	93	207
	0.551	0.449	
	0.826	0.564	
	0.376	0.307	
Column Total			303
		138	165
		0.455	0.545

Per l'anàlisi de la variable `output` en funció del `sex` observem que hi ha menys proporció de dones amb malalties relacionades amb el cor que d'homes.

```
> barplot(table(all_data$sex, all_data$output),
+         beside = T,
+         col = c("yellow", "lightblue"),
+         names = c("No", "Yes"),
+         legend.text = c("Female", "Male"))
```



La prova de Fisher parteix de la hipòtesi nul·la que les dues variables són independents, és a dir, els valors d'una no depenen dels valors de l'altra. Aquest test només ens indicarà si hi ha diferència estadísticament significativa de la variable "output" en funció del sexe, que tal i com es pot observar amb el p-value < 0.05, concloem que sí que hi ha diferències, però no sabem la força d'aquesta diferència.

```
> fisher.test(x = tabla1, alternative = "two.sided")

Fisher's Exact Test for Count Data

data:  tabla1
p-value = 1.042e-06
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1519598 0.4783553
sample estimates:
odds ratio
 0.2731136
```

Aleshores, fent servir la funció `assocstats()` podem analitzar la força de l'associació, que en aquest cas és mitjà, ni molt petita ni molt gran.

```
> assocstats(x = tabla1)
              X^2 df    P(> X^2)
Likelihood Ratio 24.841  1 6.2263e-07
Pearson          23.914  1 1.0072e-06

Phi-Coefficient   : 0.281
Contingency Coeff.: 0.27
Cramer's V        : 0.281
```

Analitzarem l'homogeneïtat de les dades amb l'ús de la funció `var.test()` per determinar si la variància és igual o diferent. Podem observar que el p-value en ser superior a 0.05, farà que haguem de rebutjar la hipòtesi nul·la d'igualtat de variàncies entre homes i dones per la variable "output".

```
> x1 <- all_data$output[all_data$sex == 0]
> x2 <- all_data$output[all_data$sex == 1]
> var.test(x1, x2)

F test to compare two variances

data:  x1 and x2
F = 0.76208, num df = 95, denom df = 206, p-value = 0.1343
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5455293 1.0885394
sample estimates:
ratio of variances
 0.7620767
```

### Anàlisi variable OUTPUT en funció de DIABETIS (fbs)

```
> # OUTPUT vs. DIABETES
> tabla2 <- table(all_data$fbs, all_data$output, dnn = c("Diabetes", "Output"))
> tabla2
      Output
Diabetis  0  1
0      116 142
1       22  23
```



```
> prop.table(table(all_data$fbns, all_data$output))
      0      1
0 0.38283828 0.46864686
1 0.07260726 0.07590759
> CrossTable(all_data$fbns, all_data$output, prop.chisq = FALSE)
```

cell contents

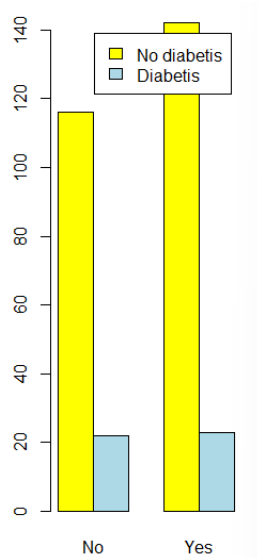
		N
N / Row Total		
N / Col Total		
N / Table Total		

Total Observations in Table: 303

all_data\$fbns	all_data\$output		
	0	1	Row Total
0	116 0.450 0.841 0.383	142 0.550 0.861 0.469	258 0.851
1	22 0.489 0.159 0.073	23 0.511 0.139 0.076	45 0.149
Column Total	138 0.455	165 0.545	303

En canvi, quan fem el mateix anàlisi per la diabetis, observem que no s'aprecien diferències notables entre els diabètics i els no diabètics en relació a les malalties relacionades amb el cor.

```
> barplot(table(all_data$fbns, all_data$output),
+         beside = T,
+         col = c("yellow", "lightblue"),
+         names = c("No", "Yes"),
+         legend.text = c("No diabetis", "Diabetis"))
```



Observem que fent la prova exacta de Fisher es pot concloure que no hi ha diferència estadísticament significativa de la variable "output" en funció de si el pacient presenta o no diabetis, perquè el p-value és > 0.05.

```
> fisher.test(x = tabla2, alternative = "two.sided")

Fisher's Exact Test for Count Data

data:  tabla2
p-value = 0.6308
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4308961 1.6975867
sample estimates:
odds ratio
0.8544825
```

També observem que la força de l'associació és molt baixa. Tot i que no caldria comprovar-ho perquè no s'han observat diferències estadísticament significatives entre aquests grups.

```
> assocstats(x = tabla2)
              X^2 df P(> X^2)
Likelihood Ratio 0.23770 1 0.62587
Pearson          0.23833 1 0.62542

Phi-Coefficient   : 0.028
Contingency Coeff.: 0.028
Cramer's V       : 0.028
```

Analitzant de nou l'homogeneïtat de les dades, podem observar que el p-value en aquest cas és també superior a 0.05. Aleshores haurem de rebutjar la hipòtesi nul·la d'igualtat de variàncies entre pacients amb diabetis i sense per la variable "output".

```
> x1 <- all_data$output[all_data$fbs == 0]
> x2 <- all_data$output[all_data$fbs == 1]
> var.test(x1, x2)

F test to compare two variances

data: x1 and x2
F = 0.97209, num df = 257, denom df = 44, p-value = 0.8593
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5931042 1.4791731
sample estimates:
ratio of variances
 0.9720938
```

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

En aquest anàlisi volem analitzar la variable "output" que ens indica si la persona ha patit alguna malaltia relacionada amb el cor en funció de totes les altres variables explicatives del fitxer "heart.csv" juntament amb el de "o2Saturation.csv".

Abans de seleccionar els grups de dades que es volen analitzar, Un cop comprovat que no hi ha cap col·linealitat, passem a realitzar el model lineal.

```
> round(cor(all_data), 3)

index      index  age      sex      cp      trtbps      chol      fbs      restecg      thalachh      exng      oldpeak      slp      caa      thall      output      o2.saturation
index      1.000  0.185  0.201 -0.399  0.109  0.020  0.001 -0.021 -0.405  0.364  0.298 -0.276  0.385  0.257 -0.863  -0.412
age         0.185  1.000 -0.098 -0.069  0.279  0.214  0.121 -0.116 -0.399  0.097  0.210 -0.169  0.276  0.068 -0.225  0.019
sex         0.201 -0.098  1.000 -0.049 -0.057 -0.198  0.045 -0.058 -0.044  0.142  0.096 -0.031  0.118  0.210 -0.281  -0.134
cp          -0.399 -0.069 -0.049  1.000  0.048 -0.077  0.094  0.044  0.296 -0.394 -0.149  0.120 -0.181 -0.162  0.434  0.136
trtbps      0.109  0.279 -0.057  0.048  1.000  0.123  0.178 -0.114 -0.047  0.068  0.193 -0.121  0.101  0.062 -0.145  0.046
chol        0.020  0.214 -0.198 -0.077  0.123  1.000  0.013 -0.151 -0.010  0.067  0.054 -0.004  0.071  0.099 -0.085  -0.030
fbs         0.001  0.121  0.045  0.094  0.178  0.013  1.000 -0.084 -0.009  0.026  0.006 -0.060  0.138 -0.032 -0.028  -0.055
restecg     -0.021 -0.116 -0.058  0.044 -0.114 -0.151 -0.084  1.000  0.044 -0.071 -0.059  0.093 -0.072 -0.012  0.137  0.070
thalachh    -0.405 -0.399 -0.044  0.296 -0.047 -0.010 -0.009  0.044  1.000 -0.379 -0.344  0.387 -0.213 -0.096  0.422  0.150
exng        0.364  0.097  0.142 -0.394  0.068  0.067  0.026 -0.071 -0.379  1.000  0.288 -0.258  0.116  0.207 -0.437  -0.081
oldpeak     0.298  0.210  0.096 -0.149  0.193  0.054  0.006 -0.059 -0.344  0.288  1.000 -0.578  0.223  0.210 -0.431  0.018
slp         -0.276 -0.169 -0.031  0.120 -0.121 -0.004 -0.060  0.093  0.387 -0.258 -0.578  1.000 -0.080 -0.105  0.346  -0.045
caa         0.385  0.276  0.118 -0.181  0.101  0.071  0.138 -0.072 -0.213  0.116  0.223 -0.080  1.000  0.152 -0.392  -0.133
thall       0.257  0.068  0.210 -0.162  0.062  0.099 -0.032 -0.012 -0.096  0.207  0.210 -0.105  0.152  1.000 -0.344  -0.011
output      -0.863 -0.225 -0.281  0.434 -0.145 -0.085 -0.028  0.137  0.422 -0.437 -0.431  0.346 -0.392 -0.344  1.000  0.309
o2.Saturation -0.412  0.019 -0.134  0.136  0.046 -0.030 -0.055  0.070  0.150 -0.081  0.018 -0.045 -0.133 -0.011  0.309  1.000
```

### Regressió lineal

```
> mod <- lm(output ~ age + sex + cp + trtbps + chol + fbs + restecg +
+           thalachh + exng + oldpeak + slp + caa + thall +
+           o2_Saturation,
+           data = all_data)
> summary(mod)
```

```
Call:
lm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
    restecg + thalachh + exng + oldpeak + slp + caa + thall +
    o2_Saturation, data = all_data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-0.96517 -0.22594  0.03098  0.24775  0.89411
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.949e+01  5.707e+00  -5.168 4.44e-07 ***
age          -2.166e-03  2.590e-03  -0.836 0.403660
sex          -1.683e-01  4.536e-02  -3.710 0.000249 ***
cp           1.038e-01  2.146e-02   4.836 2.16e-06 ***
trtbps       -2.142e-03  1.202e-03  -1.782 0.075869 .
chol         -2.228e-04  4.039e-04  -0.552 0.581529
fbs          3.534e-02  5.713e-02   0.619 0.536711
restecg      3.797e-02  3.823e-02   0.993 0.321461
thalachh     2.071e-03  1.095e-03   1.891 0.059652 .
exng        -1.458e-01  4.910e-02  -2.969 0.003239 **
oldpeak     -6.297e-02  2.193e-02  -2.872 0.004386 **
slp          9.652e-02  4.065e-02   2.374 0.018240 *
caa         -8.972e-02  2.099e-02  -4.274 2.61e-05 ***
thall       -1.257e-01  3.410e-02  -3.684 0.000274 ***
o2_Saturation 3.129e-01  5.883e-02   5.319 2.10e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3386 on 288 degrees of freedom
Multiple R-squared:  0.5607,    Adjusted R-squared:  0.5394
F-statistic: 26.26 on 14 and 288 DF, p-value: < 2.2e-16
```

```

> # Treiem del model les variables que no són significatives:
> # age, trtbps, chol, fbs, restecg, thalachh
> mod2 <- lm(output ~ sex + cp + exng + oldpeak + slp + caa + thall, data = all_data)
> summary(mod2)

Call:
lm(formula = output ~ sex + cp + exng + oldpeak + slp + caa + thall, data = all_data)

Residuals:
    Min       1Q   Median       3Q      Max
-1.05599 -0.21775  0.06441  0.25005  0.90632

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.89183    0.11247   7.929 4.58e-14 ***
sex          -0.17638    0.04587  -3.845 0.000147 ***
cp           0.12040    0.02213   5.440 1.12e-07 ***
exng        -0.18531    0.05043  -3.674 0.000283 ***
oldpeak     -0.07078    0.02287  -3.095 0.002158 **
slp          0.11038    0.04156   2.656 0.008335 **
caa         -0.11552    0.02120  -5.448 1.07e-07 ***
thall       -0.12072    0.03583  -3.369 0.000854 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3596 on 295 degrees of freedom
Multiple R-squared:  0.4924,    Adjusted R-squared:  0.4804
F-statistic: 40.89 on 7 and 295 DF,  p-value: < 2.2e-16

```

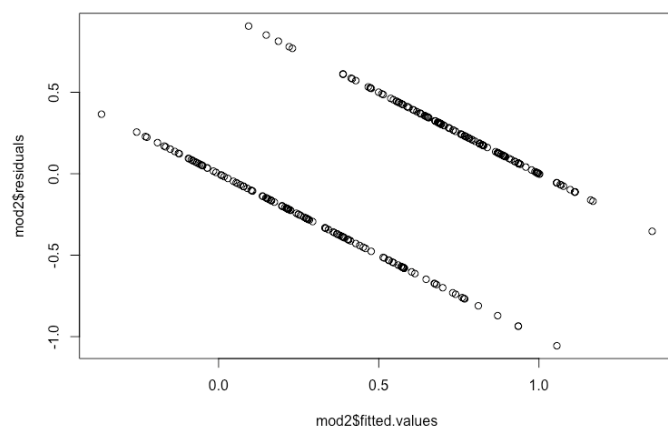
Podem observar que les variables “sex”, “cp”, “exng”, “oldpeak”, “slp”, “caa” i “thall” són les úniques que mantenim al model perquè són estadísticament significatives. Totes aquestes mostren un efecte similar sobre la variable resposta “output”.

En analitzar els residus del model podem observar que aquests presenten una estructura i que els valors tenen una tendència i no es troben de manera aleatòria al voltant del zero. Per tant, no es pot considerar el model com a correcte, perquè aquest no està ben encaixat amb les dades.

```

> # Anàlisi dels residus
> ## Gràfic de residus enfront de valors estimats
> plot(mod2$residuals ~ mod2$fitted.values)

```

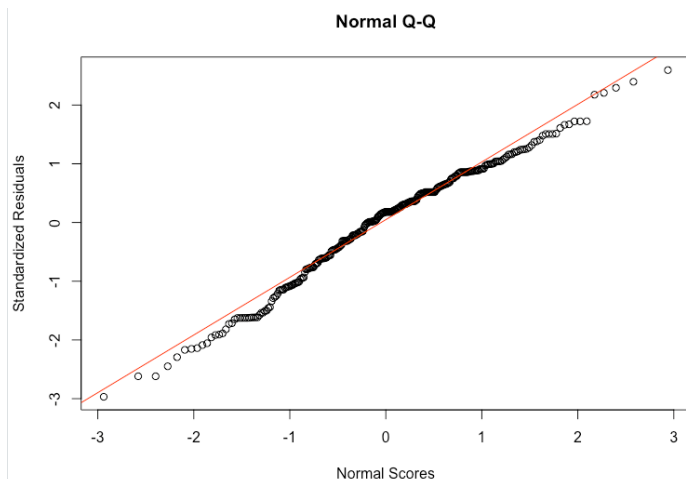


No obstant això, si que observem en el gràfic quantil-quantil que els valors s'ajusten prou bé a una distribució normal dels residus estandaritzats.

```

> #Estandaritzem els residus
> lm.stdres <- rstandard(mod2)
> #Calculem la probabilitat normal i ho comparem amb els valors dels residus estandaritzats
> qqnorm(lm.stdres,
+       ylab="Standardized Residuals",
+       xlab="Normal Scores",
+       main="Normal Q-Q")
> qqline(lm.stdres, col = "red")

```



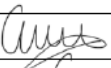

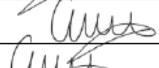

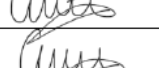



## 5. Resolució del problema

Podem concloure que el sexe si que és un factor que afecta a que els pacients presentin o no malalties relacionades amb el cor, a diferència de la diabetis on no s'han observat diferències estadísticament significatives. També quan s'ha realitzat el model lineal, s'ha vist que aquest no explica bé les dades, de manera que concloem que hi ha molts factors que poden afectar a aquestes malalties i que es difícil de predir amb les dades que s'han obtingut dels pacients.

## 6. Codi

El codi s'ha anat mostrant al llarg de cada apartat de la pràctica.

## 7. Contribucions i signatura

Contribucions	Signatura
Investigació prèvia	 
Redacció de les respostes	 
Desenvolupament del codi	 
Participació al video	 

## 8. Bibliografia

<https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>  
<https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7>  
<https://www.kaggle.com/learn/machine-learning-explainability>  
<https://www.kaggle.com/code/tentotheminus9/what-causes-heart-disease-explaining-the-model/notebook>  
<https://pubmed.ncbi.nlm.nih.gov/17488690/>  
<https://www.mayoclinic.org/diseases-conditions/heart-disease/diagnosis-treatment/drc-20353124>  
<https://www.ucsfhealth.org/education/diagnosing-heart-disease>  
<https://www.heartfoundation.org.au/bundles/your-heart/medical-tests-for-heart-disease>  
<https://www.bhf.org.uk/informationsupport/risk-factors>  
<https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack>  
[https://rpubs.com/hllinas/R\\_Barras\\_ggplot\\_unvariada](https://rpubs.com/hllinas/R_Barras_ggplot_unvariada)