TCVD Pràctica 2

Autors: Àlex López Diaz, Adrià Jaraba Currius

1. Descripció del dataset.

Perquè és important i quina pregunta/problema pretén respondre?

El dataset escollit és *Heart Attack Analysis & Prediction Dataset* penjat per *Rashik Rahman*. (https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset)

Aquest dataset conté dades referents a pacients reals amb l'objectiu de trobar alguna presencia d'alguna malaltia cardiovascular. El dataset té 303 registres i 14 columnes. Cada registre conté informació sobre la condició física del pacient, fent-lo un conjunt de dades susceptible de ser analitzat i utilitzat per predir futurs atacs de cor.

El dataset conté les següents columnes:

- age : Edat del pacient
- sex : Sexe del pacient
- cp : Tipus de dolor de pit:
    - Valor 0: angina típica
    - Valor 1: angina atípica
    - Valor 2: dolor no anginós
    - Valor 3: asimptomàtic
- trtbps: pressió arterial en repòs (in mm Hg)
- chol: colesterol en mg/dl registrar mitjançant un sensor BMI
- fbs: (sucre en sang en dejú > 120 mg/dl) (1 = true; 0 = false)
- restecg: resultats electrocardiografies en repòs
    - Valor 0: normal
    - Valor 1: tenint anomalies de l'ona ST-T
    - Valor 2: mostrant hipertròfia ventricular esquerra probable o definitiva segons els criteris d'Estes
- thalach: freqüència cardíaca màxima aconseguida
- exng: angina estable induïda per l'exercici (1 = yes; 0 = no)
- caa: vasos principals del cor (0-4)
- oldpeak - ST depression induced by exercise relative to rest
- slp: El pendent del segment ST en el punt màxim d'exercici (2 = sense pendent; 1 = plana; 0 = pendent negativa)
- thal: Un trastorn de la sang anomenat talassèmia. 1 = Sense flux sanguini en alguna part del cor; 2 = Flux normal; 3 = Hi ha flux de sang però no és normal.
- output: Variable objectiu. 0= menys probabilitat de tenir un atac de cor. 1= més probabilitat de tenir un atac de cor.

```
# Carreguem l'arxiu.
ds <- read.csv("heart.csv", sep=",")
summary(ds)
```

```
##       age             sex               cp            trtbps     
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0  
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0  
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0  
##       chol            fbs            restecg          thalachh     
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0  
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5  
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0  
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6  
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0  
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0  
##       exng           oldpeak          slp              caa        
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000  
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294  
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000  
##      thall           output      
##  Min.   :0.000   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:0.0000  
##  Median :2.000   Median :1.0000  
##  Mean   :2.314   Mean   :0.5446  
##  3rd Qu.:3.000   3rd Qu.:1.0000  
##  Max.   :3.000   Max.   :1.0000
```

```
# Tipus de variables.
str(ds)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Tots els valors són numèrics tipus “int” i “num”.

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.

Anem a estudiar les variables prescindibles del dataset aplicant un estudi de correlacions. Crearem una matriu de correlacions per trobar les columnes que més aporten a un futur anàlisi i quines podem eliminar.

És imprescindible que totes els valors de cada registre siguin numèrics, sinó no es pot aplicar aquest algorisme. Si tinguéssim valors categòrics de tipus "string" els hauríem de transformar a numèrics primer.

Les variables que menys aporten al nostre objectiu, que és predir la variable "output", son les que més s'acosten a 0. Per tant, podríem descartar les columnes "fbs" i "chol".

3. Neteja de les dades.

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

Estudiem el nombre de valors NULL i/o buits del nostre dataset. Amb la funció is.na() podem saber quants valors son nuls.

```
colSums(is.na(ds))

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

Tenim 0 valors nuls en totes les columnes.

Al tractar-se de valors numèrics, no podem trobar valors buits com a tal degut a que sempre hi haurà un número.

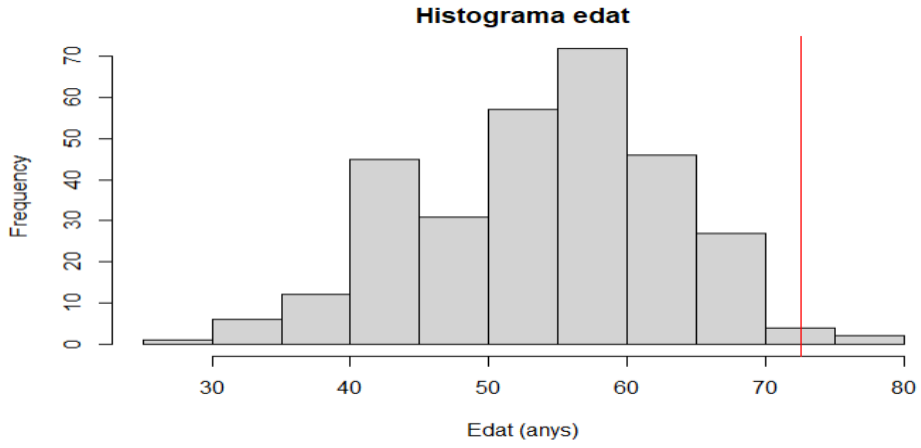
3.2. Identifica i gestiona els valors extrems.

Per trobar els valors extrems, per definició, son els que es troben a més de 2 desviacions estàndards de la mitjana. Estudiem l'edat:

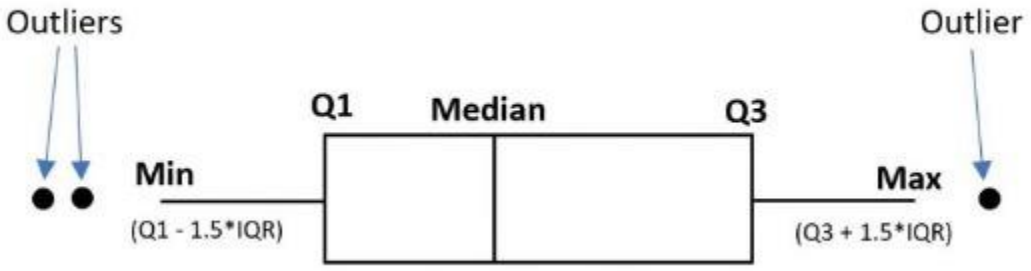
> sd_age<-mean(ds$age,na.rm=T)+2*sd(ds$age,na.rm=T)
> sd_age
[1] 72.53054

Tenint en compte que el valor màxim de l'edat és 77 anys, podem concloure que té valors extrems però no els considerarem anòmals.

Histograma de l'edat:



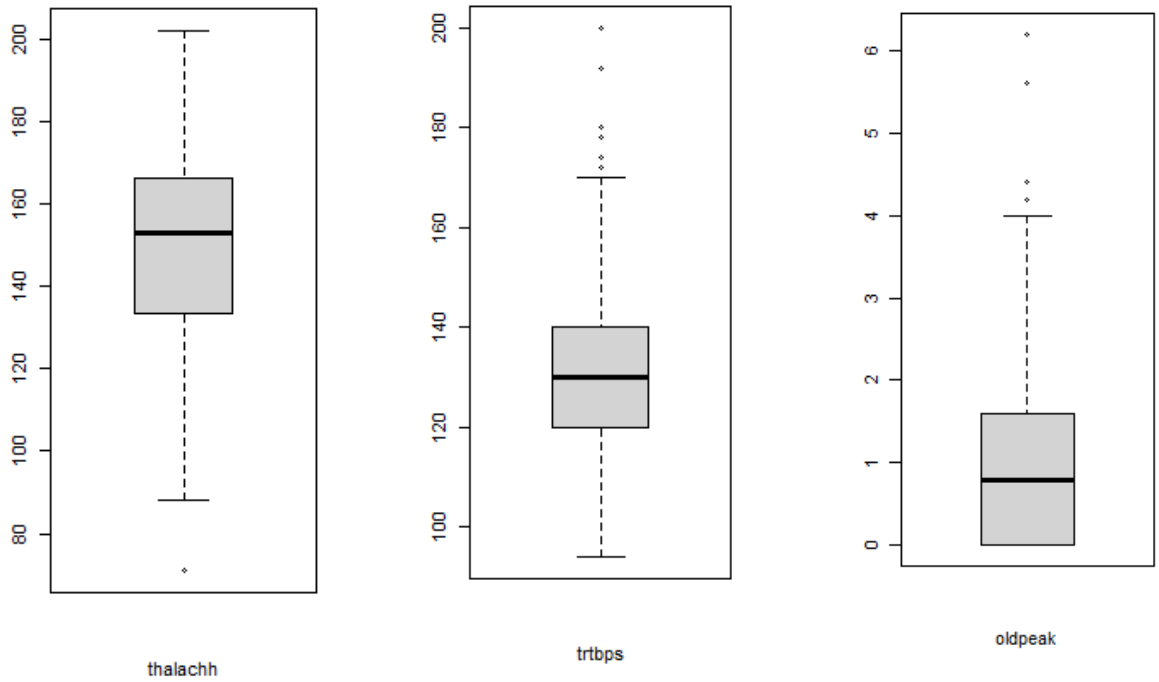
De totes formes, R disposa d'una funció anomenada boxplot() que aplica una aproximació que ens sembla més realista.

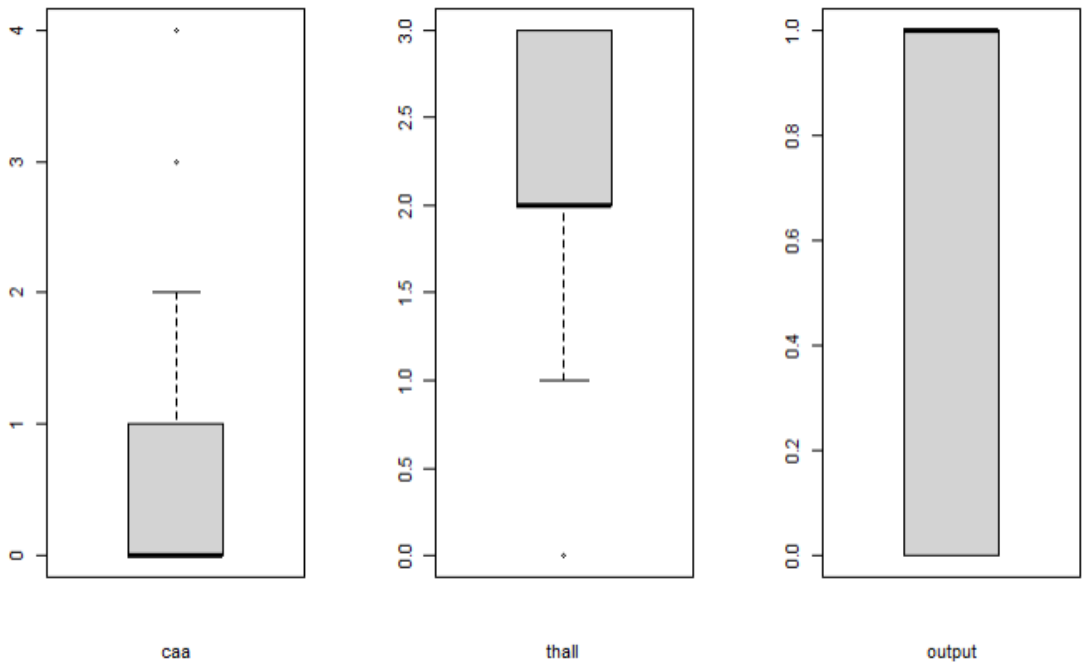


Per trobar outliers calcula la distància entre la mitjana del conjunt i els quartils Q1 i Q3. I calcula els valors màxims i mínims no outliers a través de la fórmula Q1 – 1.5 * IQR i Q3 + 1.5 * IQR on IQR és la diferència entre Q1 i Q3.

Per tant, dibuixarem tots els diagrames de caixes en busca d'aquests valors.

Diagrama de caixes

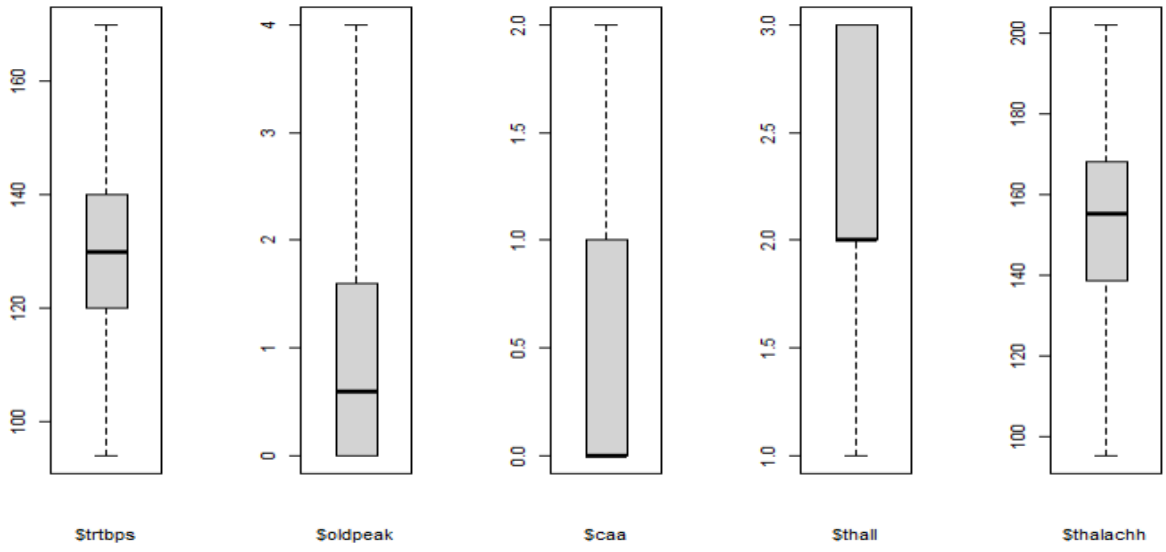




Per no ocupar molt espai hem mostrat només els boxplots amb outliers.
Boxplot es guarda els valors outliers dins una variable que es diu “out”. Per tant, per seleccionar els valors outliers és tan fàcil com:

ds <- ds[!ds$trtbps %in% boxplot.stats(ds$trtbps)$out,]
ds <- ds[!ds$oldpeak %in% boxplot.stats(ds$oldpeak)$out,]
ds <- ds[!ds$caa %in% boxplot.stats(ds$caa)$out,]
ds <- ds[!ds$thall %in% boxplot.stats(ds$thall)$out,]
ds <- ds[!ds$thalachh %in% boxplot.stats(ds$thalachh)$out,]

I el resultat final és:



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?).

Aprofitant l'estudi de correlacions realitzat per treure les columnes que menys aporten al nostre sistema, podem treure les variables que més correlacionades estan amb "output".

Si prenem un llindar, per exemple, acceptem només els atributs que tinguin una correlació absoluta major a 0.35, obtenim que les columnes més importants son:

```
result <- abs(as.data.frame(res)[c("output")])
print(filter(result, output > 0.35))
```

Description: df [6 × 1]

	output <dbl>
cp	0.4337983
thalachh	0.4217409
exng	0.4367571
oldpeak	0.4306960
caa	0.3917240
output	1.0000000

6 rows

Seleccionem les columnes que acabem de trobar:

```
ds_final$output <- factor(ds_final$output, labels=c("Risc menor","Risc major"))
ds_final$exng <- factor(ds_final$exng, labels=c("No","Si"))
ds_final$cp <- factor(ds_final$cp, labels=c("angina tipica","angina atipica","dolor no anginós","asimptomàtic"))
summary(ds_final)
```

```
       output                    cp        thalachh     exng
Risc menor:138   angina tipica   :143   Min.   : 71.0   No:204
Risc major:165   angina atipica  : 50   1st Qu.:133.5   Si: 99
                 dolor no anginós: 87   Median :153.0
                 asimptomàtic    : 23   Mean   :149.6
                                        3rd Qu.:166.0
                                        Max.   :202.0

   oldpeak          caa
Min.   :0.00   Min.   :0.0000
1st Qu.:0.00   1st Qu.:0.0000
Median :0.80   Median :0.0000
Mean   :1.04   Mean   :0.7294
3rd Qu.:1.60   3rd Qu.:1.0000
Max.   :6.20   Max.   :4.0000
```

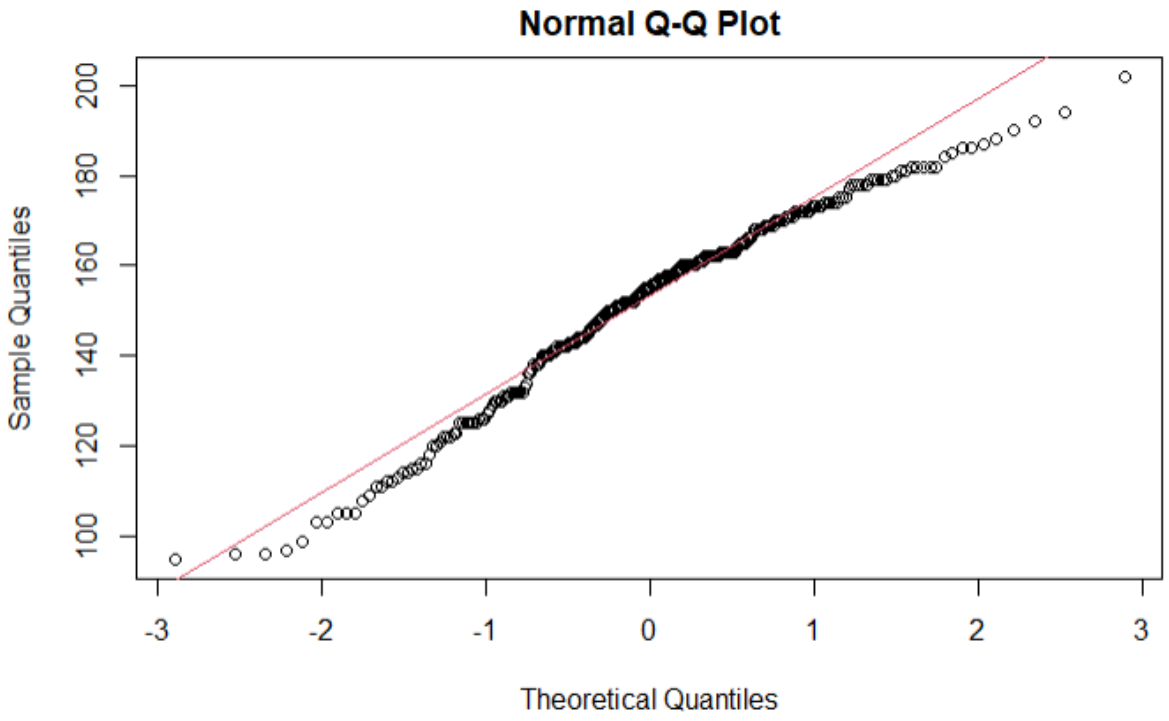
4.2. Comprovació de la normalitat i homogeneïtat de la variància.

Normalitat

Per comprovar si la distribució de les nostres variables numèriques és normal (obviem les categòriques), aplicarem el test de Shapiro-Wilk. També visualitzarem la normalitat amb la funció QQplot, que ens mostra la similitud entre les distribucions entre dos conjunts de dades i en vermell ens indica la distribució normal perfecte. Quan més s'acosti a la línia, més normal és.

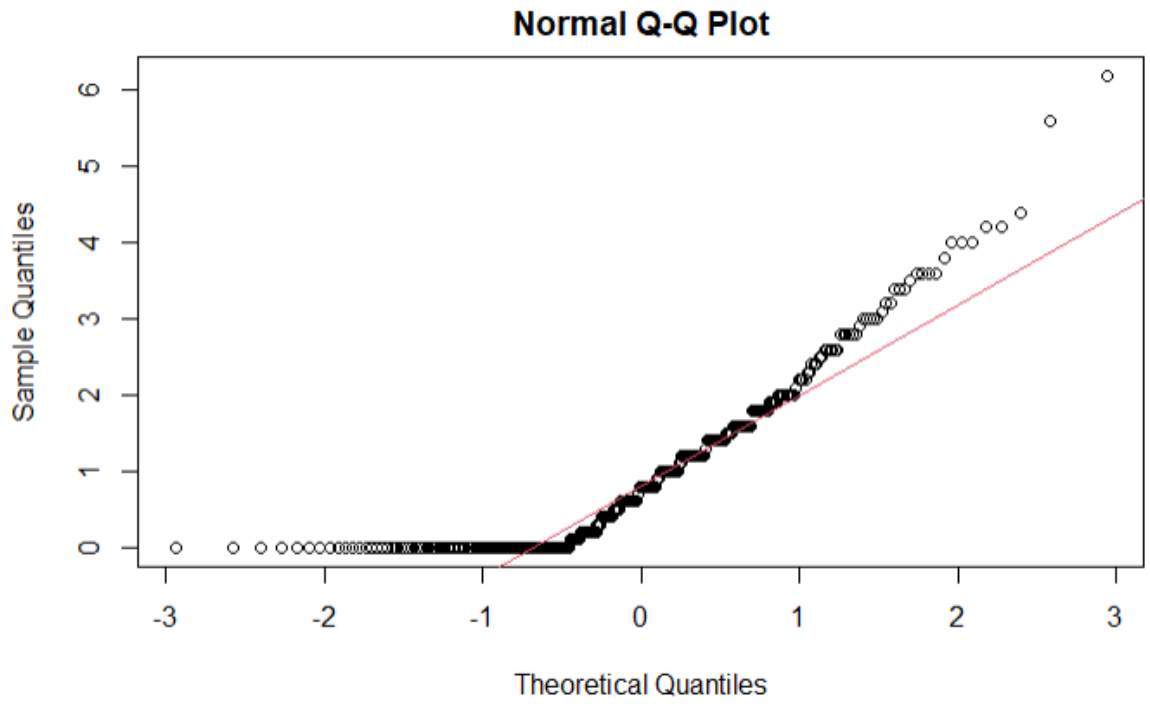
Thalachh

```
Shapiro-Wilk normality test

data:  ds$thalachh
W = 0.97327, p-value = 7.661e-05
```



Oldpeak

```
Shapiro-Wilk normality test

data:  ds$oldpeak
W = 0.85504, p-value = 5.409e-15
```



Podem afirmar que cap de les dues segueix una distribució normal al no superar un p-valor de 0.05.

Homogeneïtat de la variància

Al haver considerat que les dades no segueixen una distribució normal, si volem calcular l'homoscedasticitat de les dades aplicarem l'algoritme flinger.test().

De la mateixa manera que els algorismes per analitzar la normalitat, ens retorna un p-valor que volem que sigui > 0,05 per poder afirmar que presenten variàncies estadísticament semblants.

fligner.test(output ~ thalachh, data = ds)
fligner.test(output ~ oldpeak, data = ds)
fligner.test(output ~ caa, data = ds)
fligner.test(output ~ cp, data = ds)
fligner.test(output ~ exng, data = ds)

```
        Fligner-Killeen test of homogeneity of variances

data:  output by thalachh
Fligner-Killeen:med chi-squared = 58.787, df = 83, p-value = 0.9797
```

```
        Fligner-Killeen test of homogeneity of variances

data:  output by oldpeak
Fligner-Killeen:med chi-squared = 32.297, df = 35, p-value = 0.5993
```

```
        Fligner-Killeen test of homogeneity of variances

data:  output by caa
Fligner-Killeen:med chi-squared = 2.8846, df = 2, p-value = 0.2364
```

```
        Fligner-Killeen test of homogeneity of variances

data:  output by cp
Fligner-Killeen:med chi-squared = 7.4197, df = 3, p-value = 0.05966
```

```
        Fligner-Killeen test of homogeneity of variances

data:  output by exng
Fligner-Killeen:med chi-squared = 0.0016782, df = 1, p-value = 0.9673
```

Els resultats que hem calculat son clars, podem assegurar que la variància de “output” és similar entre totes les variables que hem escollit. La variable amb el p-valor més petit és cp, però està per sobre del llindar d’acceptació (0,05).
El test s’ha realitzat amb la variable output com a tipus int per poder realitzar els anàlisis següents.

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.

En el nostre cas, els estadístics hauran de ser no paramètrics degut que les variables triades no presenten una distribució normal.

Per a la variable dicotómica $exng apliquem el test de Wilcoxon.

wilcox.test(ds_final$output~ds_final$exng)

```
        Wilcoxon rank sum test with continuity correction

data:  ds_final$output by ds_final$exng
W = 10842, p-value = 2.293e-12
alternative hypothesis: true location shift is not equal to 0
```

Obtenim un p-value inferior a 0.05 i podem afirmar que la diferència entre els grups son significatives. La variable output serà estadísticament diferent entre afectats amb angina estable o no.

Per fer el mateix anàlisi però amb variables amb més de dues categories utilitzarem el test de Kruskal-Wallis.

kruskal.test(ds_final$output~ds_final$cp)

```
        Kruskal-Wallis rank sum test

data:  ds_final$output by ds_final$cp
Kruskal-Wallis chi-squared = 66.811, df = 3, p-value = 2.056e-14
```

Obtenim un resultat significatiu (p-value < 0.05), arribant a la mateixa conclusió anterior de que existeix diferència estadística entre els grups.

En el cas de les variables numèriques (“caa” tractada com a tal) realitzem els test per observar la correlació de Spearman.

cor.test(ds_final$output,ds_final$thalachh,method="spearman")
cor.test(ds_final$output,ds_final$oldpeak,method="spearman")
cor.test(ds_final$output,ds_final$caa,method="spearman")

```
Warning: Cannot compute exact p-value with ties
        Spearman's rank correlation rho

data:  ds_final$output and ds_final$thalachh
S = 1836793, p-value = 3.303e-11
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3941705

Warning: Cannot compute exact p-value with ties
        Spearman's rank correlation rho

data:  ds_final$output and ds_final$oldpeak
S = 4272468, p-value = 4.875e-12
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.4091884

Warning: Cannot compute exact p-value with ties
        Spearman's rank correlation rho

data:  ds_final$output and ds_final$caa
S = 4471992, p-value = 3.306e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.4749977
```

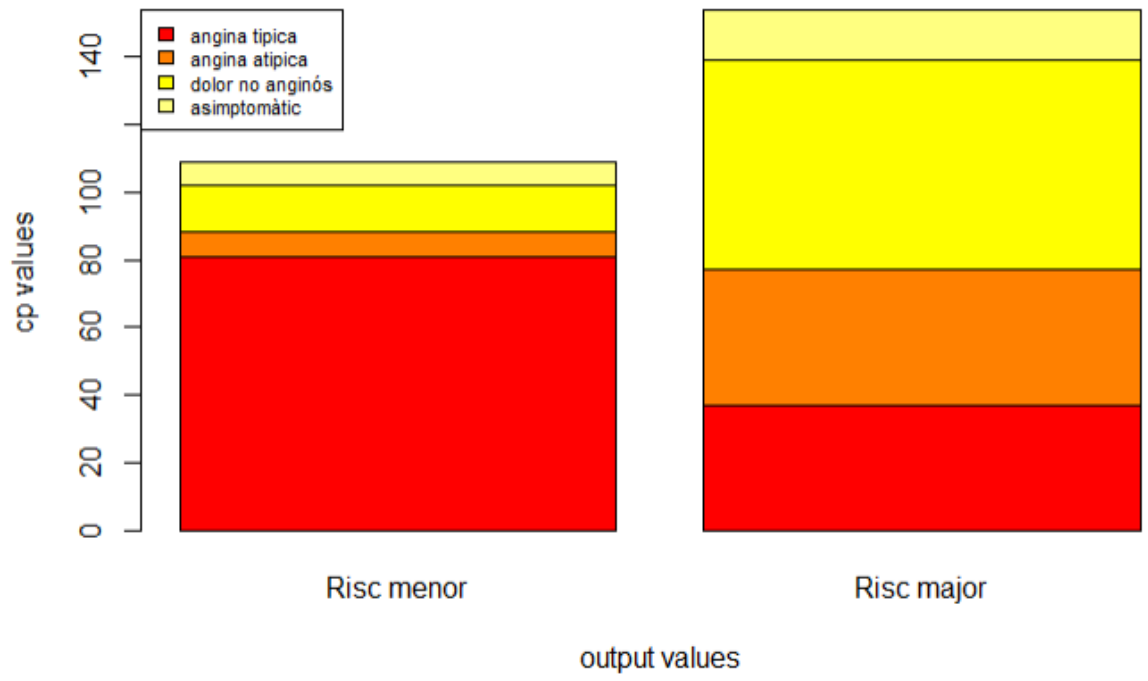
Es pot observar que la correlació és significativa (p-valor < 0.05) en les tres variables seleccionades i que prenen els següents valors:

- $thalachh = 0.394.
- $oldpeak = -0.409.
- $caa = -0.474.

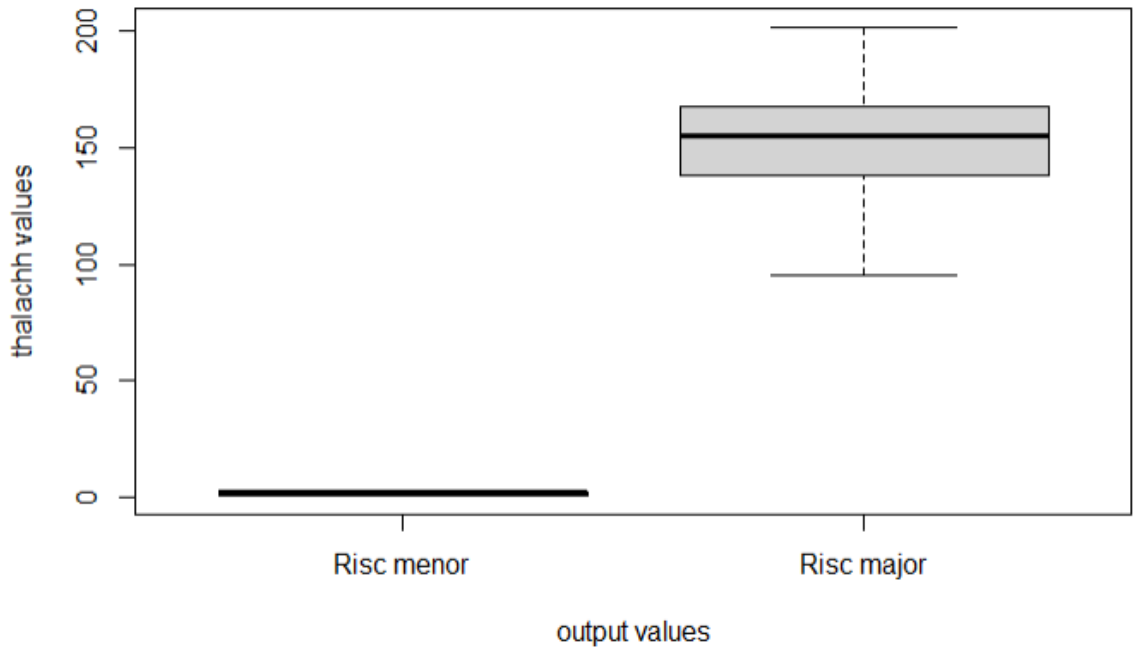
5. Representació dels resultats a partir de taules i gràfiques.

Aquest apartat es pot respondre al llarg de la pràctica, sense la necessitat de concentrar totes les representacions en aquest punt de la pràctica.

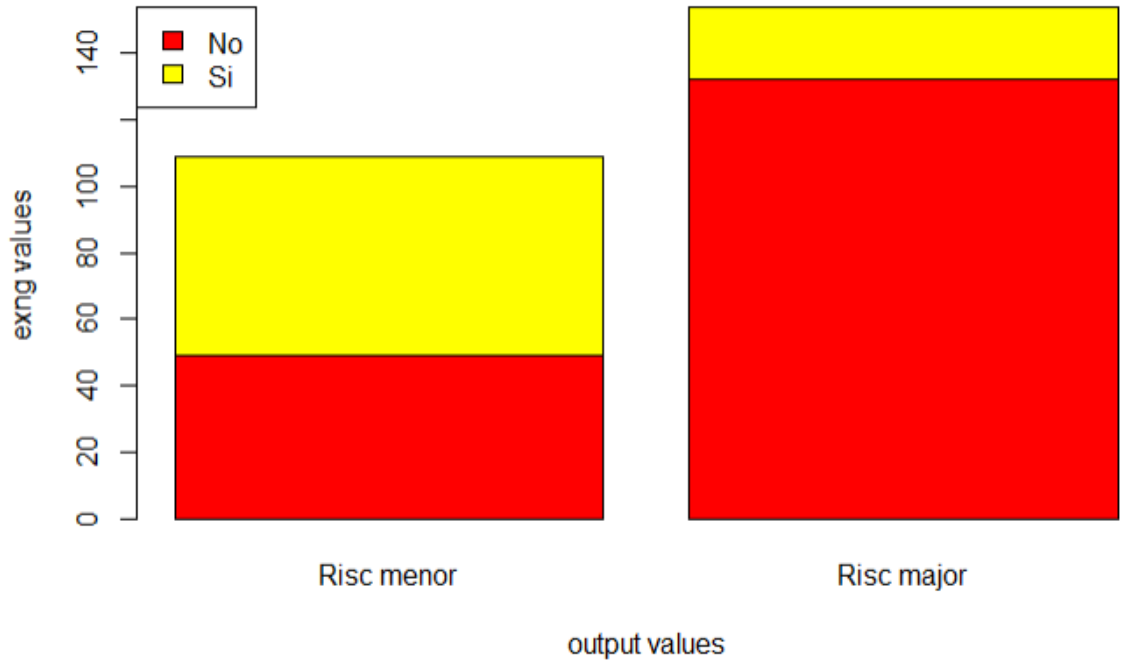
Al llarg del document es poden trobar diferents gràfics i representacions, però volem representar les variables que hem transformat i estudiat envers la variable target per tenir una representació visual de les dades finals.



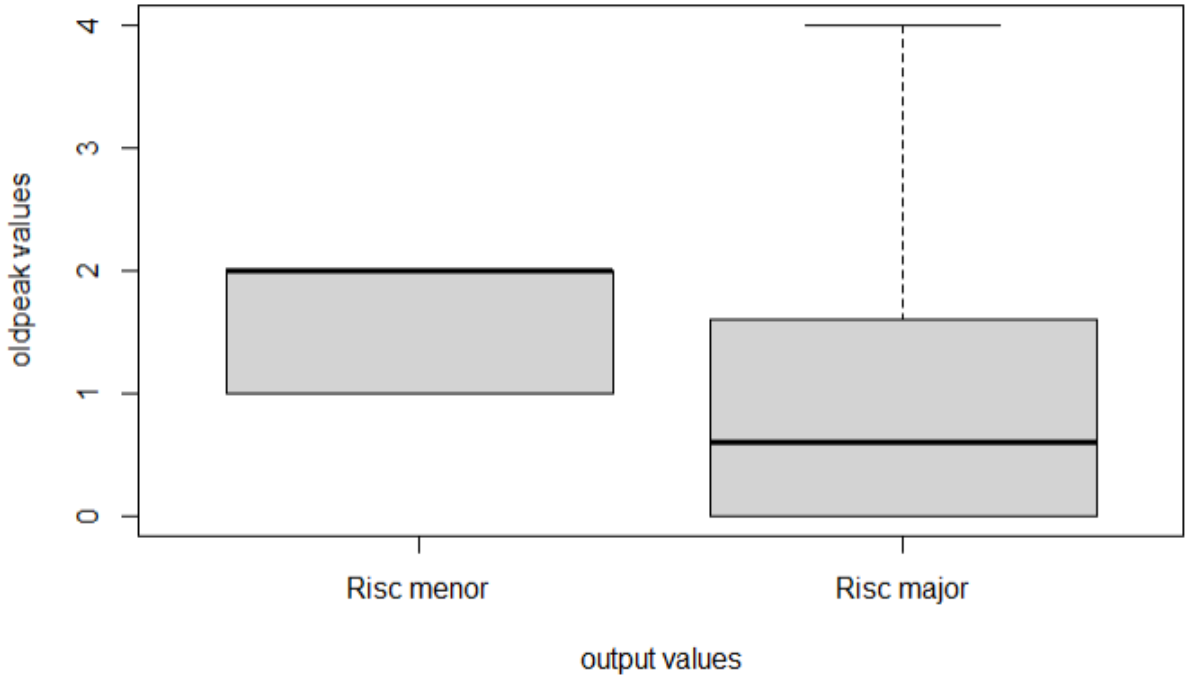
Podem apreciar que una angina típica és el cas més normal dins el risc menor, mentre que presentar angina atípica o dolor comporta un risc major.



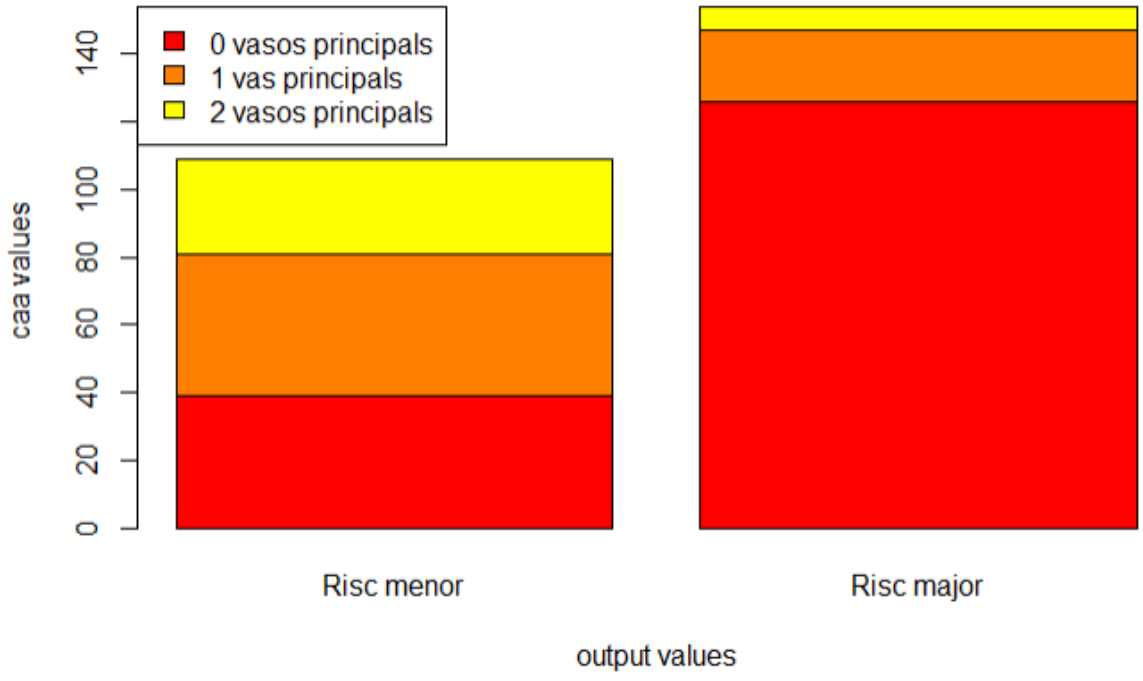
Apreciem clarament que aquestes dades son directament proporcionals, quant més alt és thalachh, major risc presenta el pacient.



Veiem representat un augment del risc en pacients que no tenen aquesta angina (o dolor de pit) produïda per l'exercici.



S'aprecia en el boxplot anterior que un valor baix del paràmetre oldpeak representa un risc major envers al que el tenen alt.



Per acabar, veiem un risc major en les persones que disposen de 0 vasos principals al seu cor envers als que en tenen més de 0.

6. Resolució del problema.

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

En quant a l'estructura de les dades, podem afirmar que disposem de 305 registres amb 14 atributs, dels quals hem escollit per analitzar els 5 més significatius per analitzar. L'objectiu de les dades és predir la variable "output" que representa el risc del pacient de patir un atac de cor.

Els atributs escollits no presenten una distribució normal aparent però presenten variàncies estadísticament similars amb la variable objectiu.

Hem aplicat diversos algorismes per comparar els grups de dades amb l'objectiu, resultant en que dos grups presenten diferències estadísticament significatives (exng i cp) mentre que els altres 3 eren similars (thalachh, oldpeak i caa)

En quant a les dades, podem afirmar que les persones amb una angina típica presenten un risc menor de patir un atac de cor, mentre que presentar angina atípica o dolor comporta un risc major. No tenir dolor no sembla ser molt significatiu.

Tenir una freqüència cardíaca elevada comporta un risc molt alt de patir un atac i tenir una freqüència més baixa, representa un risc menor.

Sembla haver-hi una relació entre tenir una angina produïda per l'exercici o dolor de pit al realitzar exercici. Els pacients que no han patit aquest dolor tenen una major probabilitat de tenir un atac. Curiós.

Un altre factor que produeix un alt risc de patir un atac és haver tingut pics baixos en el segment ST d'un electrocardiograma (oldpeak).

Quan un pacient té pocs vasos principals del cor funcionals el risc de patir un atac de cor augmenta exponencialment.

7. Codi.

Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

El codi utilitzat es R.
Link al repositori:
Alopezd24/TCVD_PRA2_Adria_Alex (github.com)

8. Vídeo.

Realitzar un breu vídeo explicatiu de la pràctica (màxim 10 minuts) on tots els integrants de l'equip expliquin amb les seves pròpies paraules el desenvolupament de la pràctica, basant-se en les preguntes de l'enunciat per a justificar i explicar el codi desenvolupat. Aquest vídeo s'haurà de lliurar a través d'un enllaç al Google Drive de la UOC (https://drive.google.com/…), juntament amb l'enllaç al repositori Git lliurat.

Link al vídeo:
https://drive.google.com/file/d/1ncgLlB64ipR1CZ_iLEit4G1PflxxWsPK/view?usp=share_link

Contribucions	Signatura
Investigació prèvia	**AL, AJ**
Redacció de les respostes	**AL, AJ**
Desenvolupament del codi	**AL, AJ**
Participació al vídeo	**AL, AJ**