

PRACTICA 2 - Tipologia i cicle de vida de les dades

Autor: Lidia Toda i Sergi Garcia

Desembre 2022

Contents

| | |
|---|-----------|
| Descripció del dataset | 1 |
| Visualització de les variables | 3 |
| Neteja de dades (valors nuls i outliers) | 10 |
| Normalitat de dades | 15 |
| Model lineal | 16 |
| Anàlisis i correlacions | 19 |

Carreguem llibreries.

```
library(ggplot2)
library(knitr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

Descripció del dataset

El dataset ens ofereix informació de 303 pacients amb 14 variables:

1. age : Age of the patient
2. sex : Sex of the patient
3. cp : Chest Pain type chest pain type Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic
4. trtbps : resting blood pressure (in mm Hg)
5. chol : cholestoral in mg/dl fetched via BMI sensor
6. fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. 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
8. thalach : maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak: previous peak
11. slp: slope
12. caa: number of major vessels (0-3)
13. thall: thal rate
14. output : 0= less chance of heart attack 1= more chance of heart attack

El nostre objectiu serà realitzar la neteja de dades per establir visualitzacions que ens permetin interpretar de manera fàcil el contingut i obtenir informació de quines són les variables més influents alhora de desenvolupar un atac de cor. De la mateixa manera, intentarem realitzar prediccions sobre noves incorporacions de dades.

Llegim el dataset

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

Imprimim les primeres línies del data

```
head(data)
```

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

Obtenim informació de rows i columns.

```
nombre_rows <- nrow(data)
nombre_columnes <- ncol(data)
cat("El nombre de files és de", nombre_rows, "i el nombre de columnes és de", nombre_columnes)
```

```
## El nombre de files és de 303 i el nombre de columnes és de 14
```

Obtenim informació bàsica de les variables

```
summary(data)
```

```
##      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
```

Llegim la tipologia de data de les variables

```
variables <- sapply(data,class)
kable(data.frame(variables=names(variables),clase=as.vector(variables)))
```

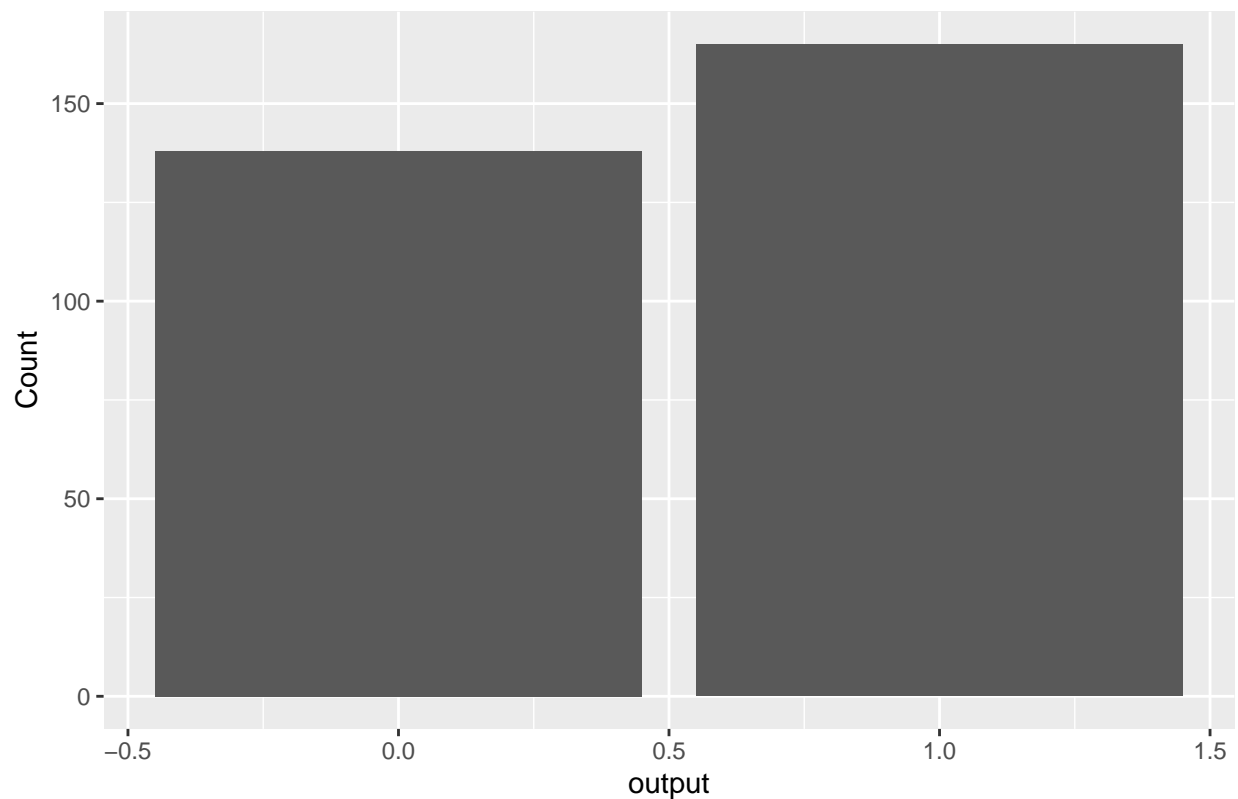
| variables | clase |
|-----------|---------|
| age | integer |
| sex | integer |
| cp | integer |
| trtbps | integer |
| chol | integer |
| fbs | integer |
| restecg | integer |
| thalachh | integer |
| exng | integer |
| oldpeak | numeric |
| slp | integer |
| caa | integer |
| thall | integer |
| output | integer |

Visualització de les variables

Observem la distribució de la columna output

```
output_plot <-ggplot(data,aes(output)) + geom_bar() + labs(x="output", y="Count") + guides(fill=guide_l  
output_plot
```

Distribució de output del dataset



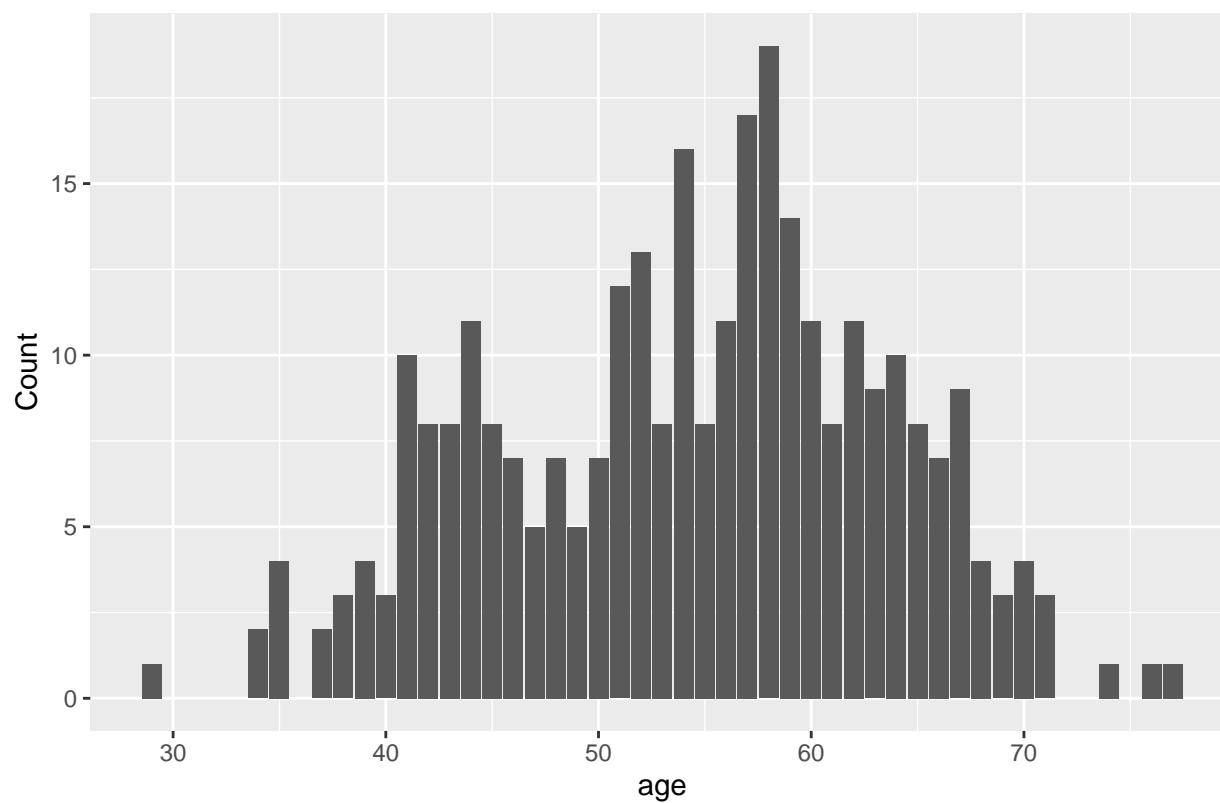
```
output0 <- sum(data$output == 0)
output1 <- sum(data$output == 1)
percentatge_output0 = (output0/nombre_rows)*100
percentatge_output1 = (output1/nombre_rows)*100
cat("El percentatge de pacients amb output positiu és de", percentatge_output1, "mentre que el percentatge de pacients amb output negatiu és de", percentatge_output0, "%\n")
```

El percentatge de pacients amb output positiu és de 54.45545 mentre que el percentatge de pacients amb output negatiu és de 45.54455

Observem la distribució de la columna age

```
age_plot <- ggplot(data, aes(age)) + geom_bar() + labs(x="age", y="Count") + guides(fill=guide_legend(title=""))
age_plot
```

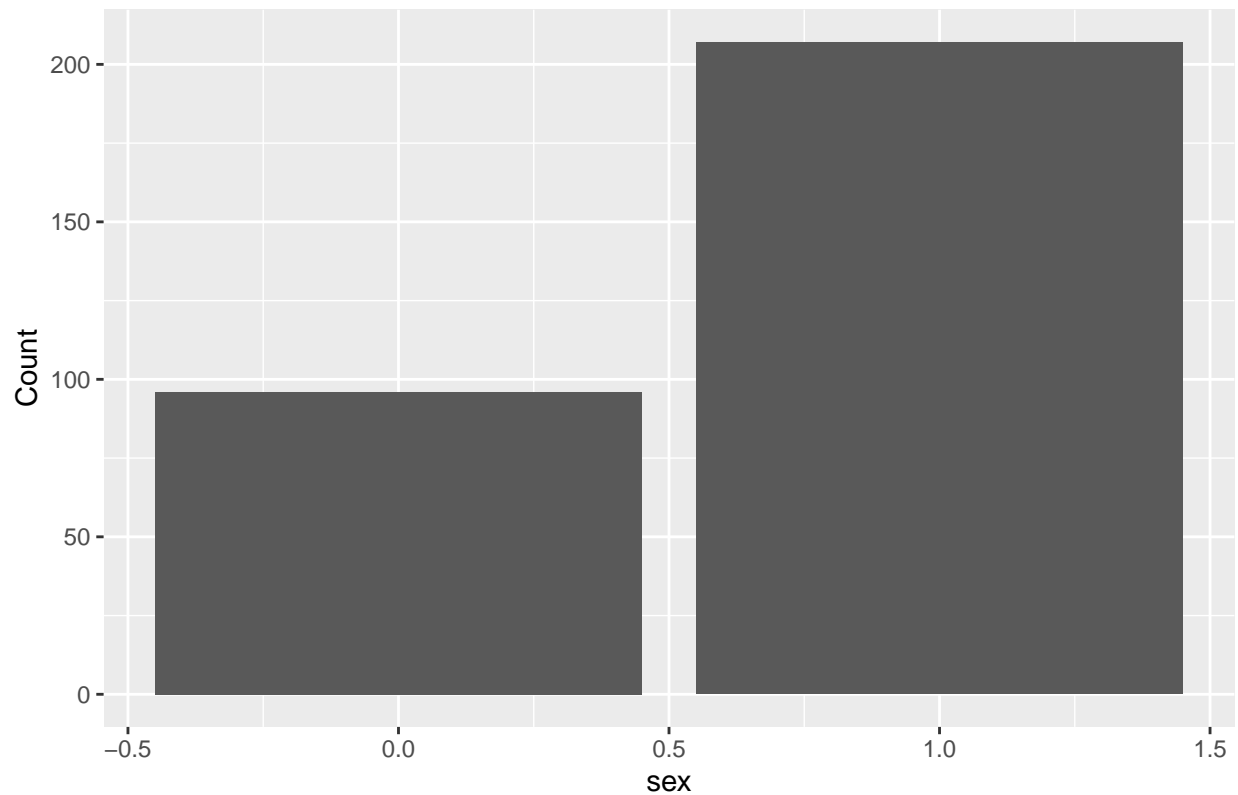
Distribució de age del dataset



Observem la distribució de la columna sex

```
sex_plot <-ggplot(data,aes(sex)) + geom_bar() + labs(x="sex", y="Count") + guides(fill=guide_legend(tit
sex_plot
```

Distribució de sex del dataset



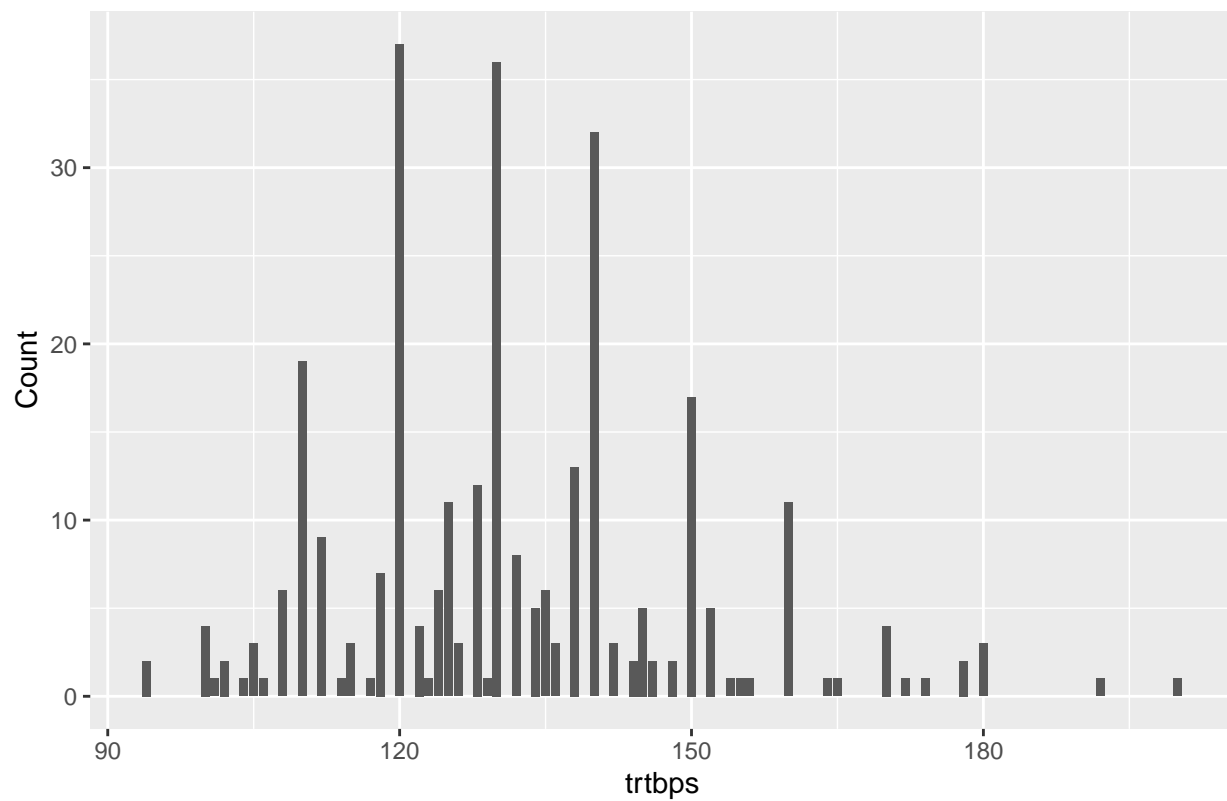
```
sex0 <- sum(data$sex == 0)
sex1 <- sum(data$sex == 1)
percentatge_sex0 = (sex0/nombre_rows)*100
percentatge_sex1 = (sex1/nombre_rows)*100
cat("El percentatge de pacients dones Ãs de", percentatge_sex1, "mentre que el percentatge de pacients
```

El percentatge de pacients dones Ãs de 68.31683 mentre que el percentatge de pacients amb output hor

Observem la distribuci3 de la columna trtbps

```
trtbps_plot <-ggplot(data,aes(trtbps)) + geom_bar() + labs(x="trtbps", y="Count") + guides(fill=guide_l
trtbps_plot
```

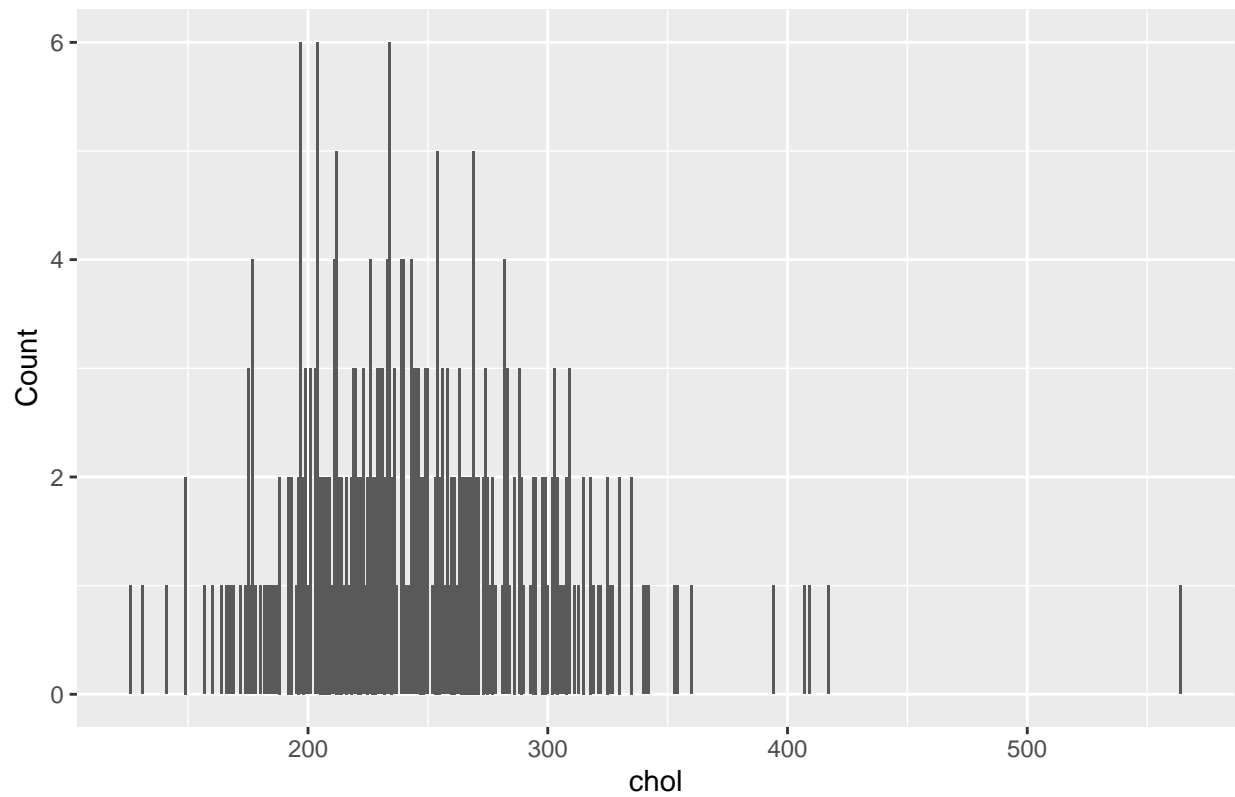
Distribució de trtbps del dataset



Observem la distribució de la columna chol

```
chol_plot <-ggplot(data,aes(chol)) + geom_bar() + labs(x="chol", y="Count") + guides(fill=guide_legend())  
chol_plot
```

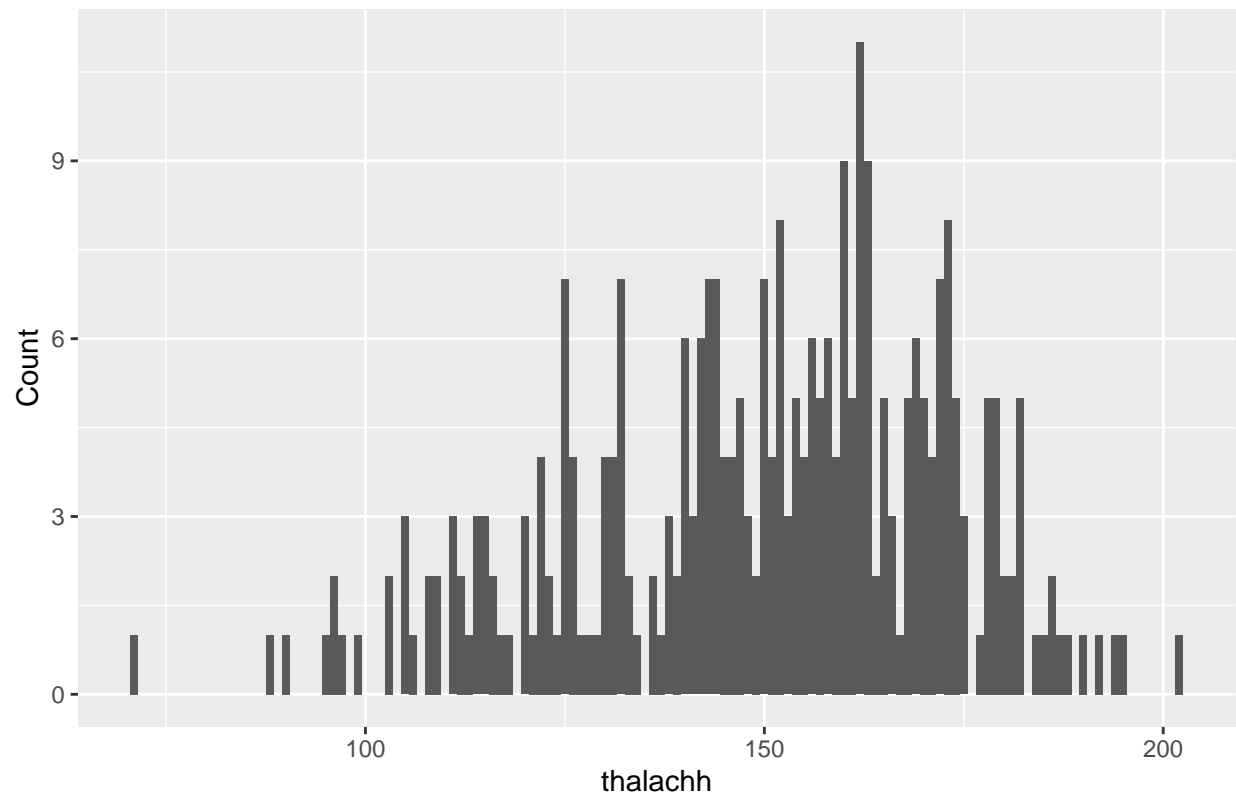
Distribució de chol del dataset



Observem la distribució de la columna thalachh

```
thalachh_plot <-ggplot(data,aes(thalachh)) + geom_bar() + labs(x="thalachh", y="Count") + guides(fill=g  
thalachh_plot
```

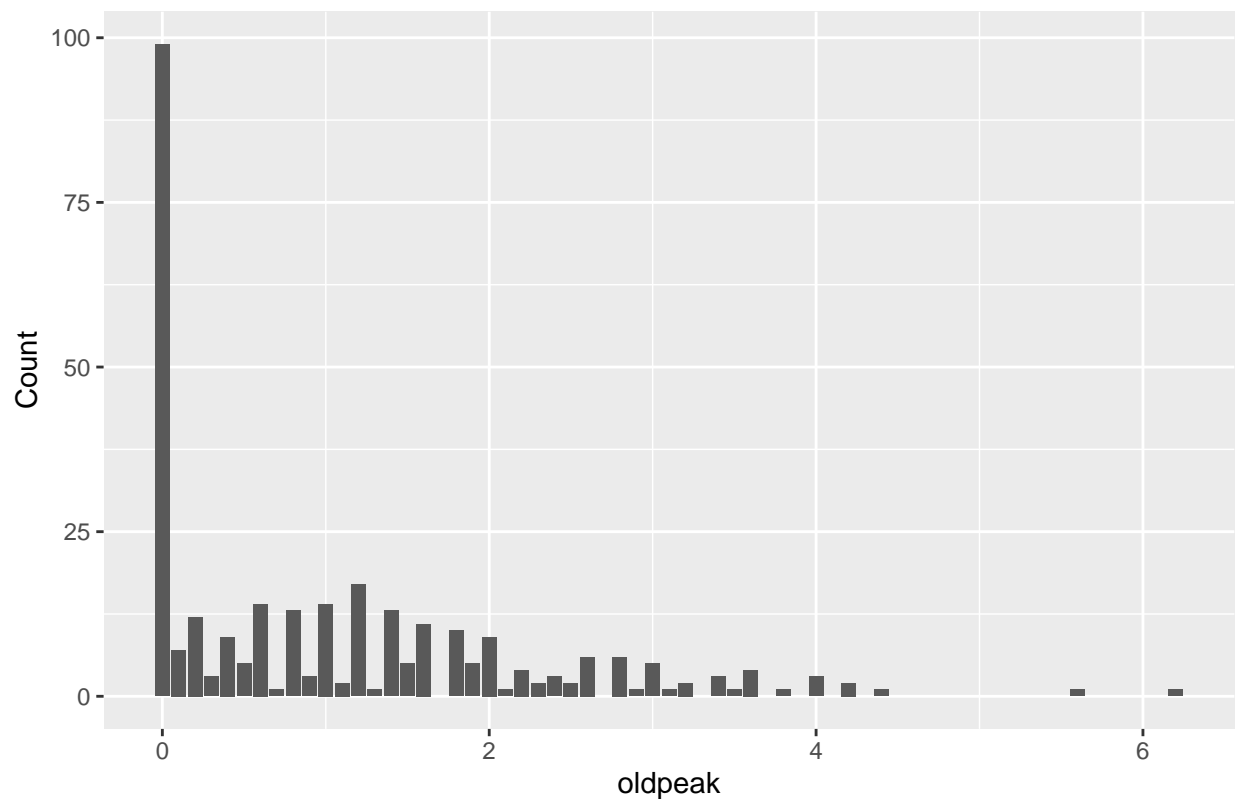

Distribució de thalachh del dataset



Observem la distribució de la columna oldpeak

```
oldpeak_plot <-ggplot(data,aes(oldpeak)) + geom_bar() + labs(x="oldpeak", y="Count") + guides(fill=guides(fill="oldpeak"))
oldpeak_plot
```

Distribució de oldpeak del dataset



Neteja de dades (valors nuls i outliers)

Comprovem si hi han valors “na” en el doc.

```
sum(is.na(data))
```

```
## [1] 0
```

Comprovem si existeixen valors duplicats al dataset

```
duplicats <- duplicated(data)
print(sum(duplicats))
```

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

Eliminem files duplicades

```
data <- unique(data)
```

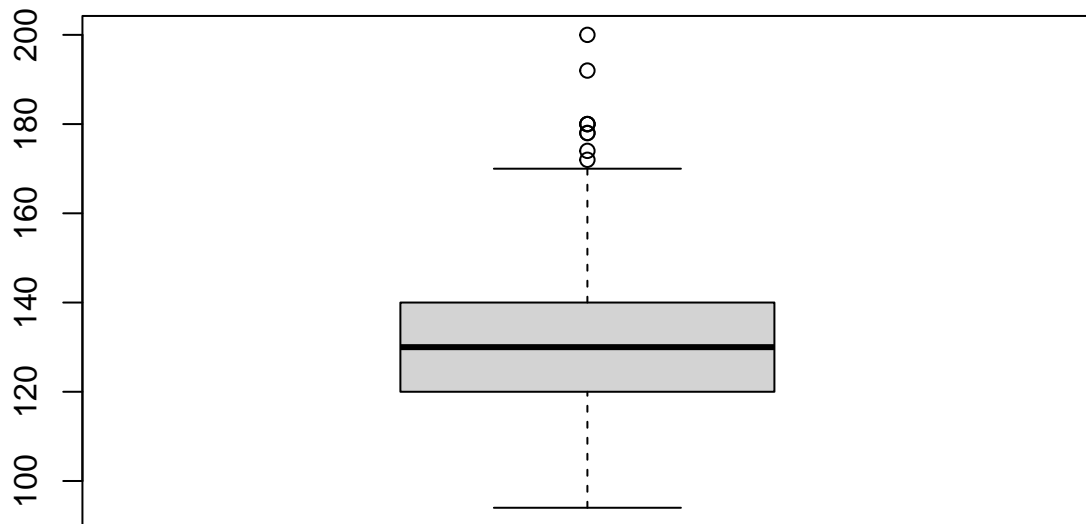
Obtenim informació de les noves rows i columnes.

```
nombre_rows <- nrow(data)
nombre_columnes <- ncol(data)
cat("El nombre de files és de", nombre_rows, "i el nombre de columnes és de", nombre_columnes)
```

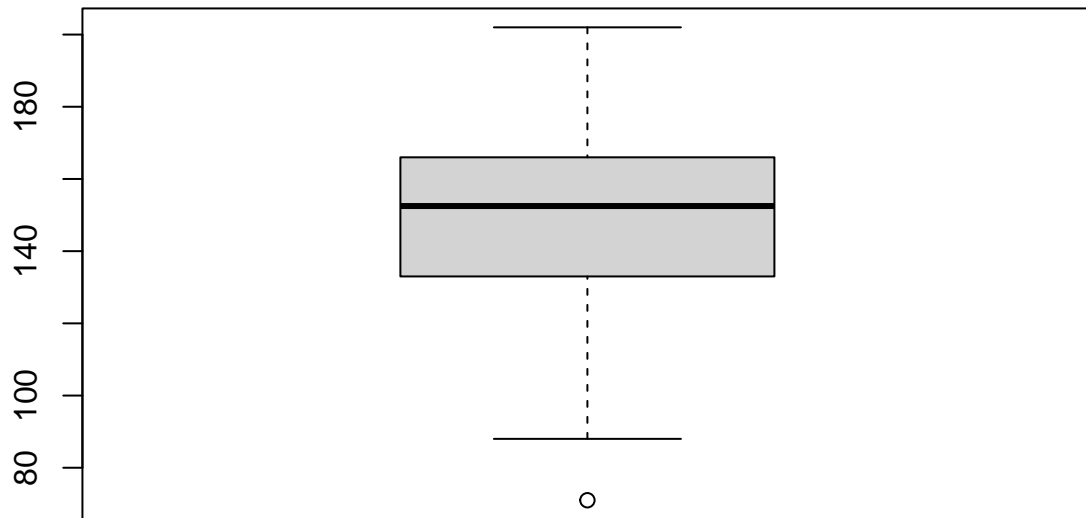
```
## El nombre de files és de 302 i el nombre de columnes és de 14
```

Busquem valors outliers en les columnes trtbps, thalachh, chol i oldpeak.

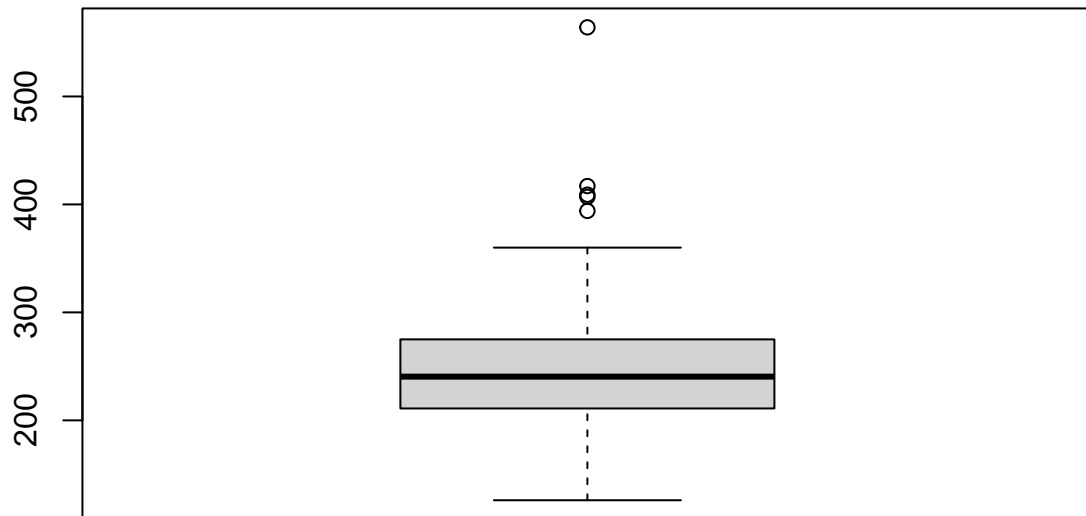
```
trtbps <- boxplot(data$trtbps)
```



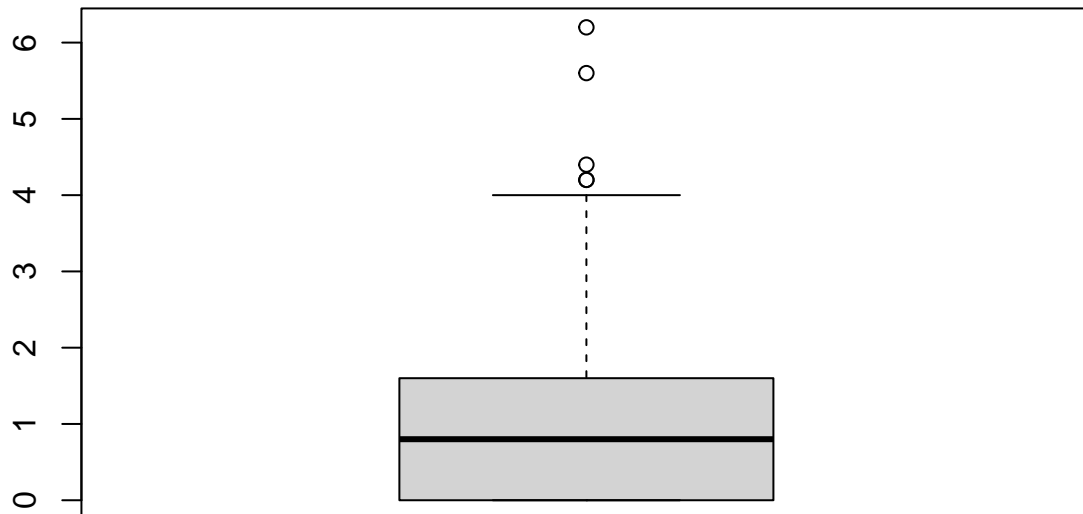
```
thalachh <- boxplot(data$thalachh)
```



```
chol <- boxplot(data$chol)
```



```
oldpeak <- boxplot(data$oldpeak)
```



```
trtbps170 <- sum(data$trtbps >= 170)
thalachh90 <- sum(data$thalachh <= 90)
chol350 <- sum(data$chol >= 350)
oldpeak_4 <- sum(data$oldpeak >= 4)
cat("La suma de files amb outliers a la columna trtbps és de", trtbps170, "\n")
```

```
## La suma de files amb outliers a la columna trtbps és de 13
```

```
cat("La suma de les files amb outliers a la columna thalachh és de", thalachh90, "\n")
```

```
## La suma de les files amb outliers a la columna thalachh és de 3
```

```
cat("La suma de les files amb outliers a la columna chol és de", chol350, "\n")
```

```
## La suma de les files amb outliers a la columna chol és de 8
```

```
cat("La suma de les files amb outliers a la columna oldpeak és de", oldpeak_4, "\n")
```

```
## La suma de les files amb outliers a la columna oldpeak és de 8
```

Eliminem els outliers del data

```
data <- subset(data, trtbps <= 170)
data <- subset(data, thalachh >= 90)
data <- subset(data, chol <= 350)
data <- subset(data, oldpeak <= 4)
```

```
nombre_rows <- nrow(data)
nombre_columnes <- ncol(data)
cat("El nombre de files és de", nombre_rows, "i el nombre de columnes és de", nombre_columnes)
```

El nombre de files és de 279 i el nombre de columnes és de 14

Imprimim una taula de correlacions

```
taula_correlacions <- round(cor(data), 2)
print(taula_correlacions)
```

```
##      age  sex  cp trtbps  chol  fbs restecg thalachh  exng oldpeak
## age      1.00 -0.06 -0.06  0.28  0.16  0.11  -0.11  -0.42  0.09  0.21
## sex     -0.06  1.00 -0.09  0.01 -0.11  0.06  -0.09  -0.03  0.18  0.16
## cp      -0.06 -0.09  1.00  0.08 -0.07  0.08   0.10   0.28 -0.38 -0.12
## trtbps   0.28  0.01  0.08  1.00  0.10  0.13  -0.14  -0.06  0.00  0.15
## chol     0.16 -0.11 -0.07  0.10  1.00  0.03  -0.16  -0.01  0.06 -0.01
## fbs      0.11  0.06  0.08  0.13  0.03  1.00  -0.08  -0.03  0.01  0.02
## restecg  -0.11 -0.09  0.10 -0.14 -0.16 -0.08   1.00   0.10 -0.12 -0.09
## thalachh -0.42 -0.03  0.28 -0.06 -0.01 -0.03   0.10   1.00 -0.38 -0.34
## exng     0.09  0.18 -0.38  0.00  0.06  0.01  -0.12  -0.38  1.00  0.32
## oldpeak  0.21  0.16 -0.12  0.15 -0.01  0.02  -0.09  -0.34  0.32  1.00
## slp     -0.15 -0.05  0.09 -0.08  0.03 -0.07   0.12   0.37 -0.26 -0.53
## caa      0.33  0.14 -0.17  0.11  0.09  0.16  -0.09  -0.25  0.13  0.18
## thall    0.06  0.24 -0.17 -0.02  0.09 -0.06   0.03  -0.10  0.20  0.19
## output  -0.23 -0.31  0.41 -0.12 -0.11 -0.03   0.18   0.42 -0.43 -0.43
##      slp  caa thall output
## age     -0.15  0.33  0.06 -0.23
## sex     -0.05  0.14  0.24 -0.31
## cp       0.09 -0.17 -0.17  0.41
## trtbps  -0.08  0.11 -0.02 -0.12
## chol     0.03  0.09  0.09 -0.11
## fbs     -0.07  0.16 -0.06 -0.03
## restecg  0.12 -0.09  0.03  0.18
## thalachh 0.37 -0.25 -0.10  0.42
## exng    -0.26  0.13  0.20 -0.43
## oldpeak -0.53  0.18  0.19 -0.43
## slp      1.00 -0.05 -0.08  0.32
## caa     -0.05  1.00  0.15 -0.39
## thall   -0.08  0.15  1.00 -0.34
## output   0.32 -0.39 -0.34  1.00
```

Podem observar que la major correlació negativa amb output és oldpeak i la major correlació positiva cp i thalachh.

Normalitat de dades

```
shapiro_age <- shapiro.test(data$age)
shapiro_trtbps <- shapiro.test(data$trtbps)
shapiro_chol <- shapiro.test(data$chol)
shapiro_thalachh <- shapiro.test(data$thalachh)
shapiro_oldpeak <- shapiro.test(data$oldpeak)
print(shapiro_age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$age
## W = 0.98799, p-value = 0.0203
```

```

print(shapiro_trtbps)

##
## Shapiro-Wilk normality test
##
## data: data$trtbps
## W = 0.9842, p-value = 0.003591

print(shapiro_chol)

##
## Shapiro-Wilk normality test
##
## data: data$chol
## W = 0.99218, p-value = 0.1485

print(shapiro_thalachh)

##
## Shapiro-Wilk normality test
##
## data: data$thalachh
## W = 0.97474, p-value = 7.648e-05

print(shapiro_oldpeak)

##
## Shapiro-Wilk normality test
##
## data: data$oldpeak
## W = 0.85352, p-value = 1.42e-15

taula <- matrix(c("0.0203", "0.003591", "0.1485", "7.648e-05", "1.42e-15"),ncol=5,byrow=TRUE)
colnames(taula) <- c("age","trtbps", "chol", "thalachh", "oldpeak")
rownames(taula) <- c("pvalue")
taula <- as.table(taula)
taula <- kable(taula)
taula

```

| | age | trtbps | chol | thalachh | oldpeak |
|--------|--------|----------|--------|-----------|----------|
| pvalue | 0.0203 | 0.003591 | 0.1485 | 7.648e-05 | 1.42e-15 |

Veiem com la variable chol manté normalitat de dades, les demés no. Per altra banda, s'ha de tenir en compte el teorema del límit central que demostra que les mitjanes de mostres suficientment grans segueixen una distribució gairebé normal malgrat que la distribució de la població no sigui normal, i que a major mida de les mostres, la distribució s'aproxima més a una distribució normal.

Model lineal

```

ntrain <- nrow(data)*0.8
ntest <- nrow(data)*0.2
set.seed(1)
index_train<-sample(1:nrow(data),size = ntrain)
train<-data[index_train,]

```



```
test<-data[-index_train,]
model <- lm(output ~ age + sex + cp + trtbps + chol + fbs + restecg + thalachh + exng + oldpeak + slp +
summary(model)
```

```
##
## Call:
## lm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
##     restecg + thalachh + exng + oldpeak + slp + caa + thall,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83129 -0.21817  0.04336  0.24405  0.90542
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.007e+00  3.632e-01   2.772 0.006068 **
## age          7.806e-06  3.085e-03   0.003 0.997983
## sex         -1.874e-01  5.595e-02  -3.349 0.000961 ***
## cp           1.017e-01  2.679e-02   3.797 0.000192 ***
## trtbps       -2.015e-03  1.654e-03  -1.218 0.224448
## chol        -7.377e-04  5.607e-04  -1.316 0.189681
## fbs          2.349e-02  6.551e-02   0.359 0.720271
## restecg      2.510e-02  4.778e-02   0.525 0.599895
## thalachh     2.555e-03  1.356e-03   1.884 0.060890 .
## exng        -1.399e-01  6.156e-02  -2.273 0.024069 *
## oldpeak     -8.203e-02  2.926e-02  -2.803 0.005537 **
## slp          7.814e-02  4.810e-02   1.624 0.105780
## caa         -1.160e-01  2.612e-02  -4.443 1.44e-05 ***
## thall       -1.242e-01  4.219e-02  -2.944 0.003602 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3519 on 209 degrees of freedom
## Multiple R-squared:  0.5308, Adjusted R-squared:  0.5016
## F-statistic: 18.18 on 13 and 209 DF,  p-value: < 2.2e-16
```

Ens quedem nomès amb les que tenen coeficients significatius

```
model1 <- lm(output ~ sex + cp + thalachh + exng + oldpeak + caa + thall, data=train)
summary(model1)
```

```
##
## Call:
## lm(formula = output ~ sex + cp + thalachh + exng + oldpeak +
##     caa + thall, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91006 -0.21219  0.05871  0.22471  0.88323
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.638746  0.211677   3.018 0.002856 **
## sex         -0.183931  0.054568  -3.371 0.000889 ***
```

```
## cp          0.099021    0.026376    3.754 0.000224 ***
## thalachh    0.003082    0.001225    2.516 0.012610 *
## exng        -0.153394    0.061046   -2.513 0.012714 *
## oldpeak     -0.105318    0.026046   -4.044 7.34e-05 ***
## caa         -0.117385    0.024945   -4.706 4.53e-06 ***
## thall       -0.123255    0.041806   -2.948 0.003549 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3529 on 215 degrees of freedom
## Multiple R-squared:  0.5147, Adjusted R-squared:  0.4989
## F-statistic: 32.57 on 7 and 215 DF,  p-value: < 2.2e-16
```

Veiem mitjançant la funció `lm` que les variables amb major explicació sobre la variable output son sex, cp i caa. El model no és massa bo donat el Rsquared de 0.48. No obstant això intentarem predir noves incorporacions

```
prob_output <- predict(model1, test, type="response")
mc_sl<-data.frame(
  real=test$output,
  predicted= ifelse(prob_output>0.5, 1, 0),
  dif=ifelse(prob_output>0.5&test$output==1, "No", ifelse(prob_output<0.5&test$output==0, "No", "Si"))
)
colnames(mc_sl)<-c("Real", "PredicciÃ³", "Diferencia?")
kable(mc_sl)
```

| | Real | PredicciÃ³ | Diferencia? |
|-----|------|------------|-------------|
| 3 | 1 | 1 | No |
| 10 | 1 | 1 | No |
| 13 | 1 | 1 | No |
| 20 | 1 | 1 | No |
| 23 | 1 | 1 | No |
| 50 | 1 | 1 | No |
| 51 | 1 | 1 | No |
| 56 | 1 | 1 | No |
| 58 | 1 | 1 | No |
| 59 | 1 | 1 | No |
| 62 | 1 | 1 | No |
| 63 | 1 | 1 | No |
| 67 | 1 | 1 | No |
| 70 | 1 | 1 | No |
| 72 | 1 | 0 | Si |
| 80 | 1 | 1 | No |
| 87 | 1 | 1 | No |
| 100 | 1 | 1 | No |
| 104 | 1 | 1 | No |
| 117 | 1 | 1 | No |
| 120 | 1 | 1 | No |
| 122 | 1 | 1 | No |
| 127 | 1 | 1 | No |
| 133 | 1 | 1 | No |
| 134 | 1 | 1 | No |
| 135 | 1 | 1 | No |
| 136 | 1 | 1 | No |
| 139 | 1 | 0 | Si |

| | Real | Predicció | Diferència? |
|-----|------|-----------|-------------|
| 148 | 1 | 1 | No |
| 150 | 1 | 1 | No |
| 156 | 1 | 1 | No |
| 159 | 1 | 0 | Si |
| 162 | 1 | 1 | No |
| 168 | 0 | 0 | No |
| 183 | 0 | 1 | Si |
| 188 | 0 | 0 | No |
| 190 | 0 | 1 | Si |
| 193 | 0 | 0 | No |
| 201 | 0 | 1 | Si |
| 202 | 0 | 0 | No |
| 207 | 0 | 0 | No |
| 208 | 0 | 0 | No |
| 209 | 0 | 0 | No |
| 217 | 0 | 1 | Si |
| 219 | 0 | 0 | No |
| 227 | 0 | 0 | No |
| 229 | 0 | 1 | Si |
| 231 | 0 | 1 | Si |
| 239 | 0 | 0 | No |
| 260 | 0 | 0 | No |
| 265 | 0 | 0 | No |
| 266 | 0 | 0 | No |
| 274 | 0 | 0 | No |
| 275 | 0 | 0 | No |
| 282 | 0 | 1 | Si |
| 296 | 0 | 0 | No |

```
count <- sum(mc_sl$`Diferència?`=="No")
total <- nrow(mc_sl)
resultat <- (count/total)
resultat <- round(resultat,2)
cat("El total de resultats correctes ha estat del", resultat*100, "%")
```

El total de resultats correctes ha estat del 82 %

Anàlisi i correlacions

```
str(data)
```

```
## 'data.frame': 279 obs. of 14 variables:
## $ age : int 63 37 41 56 57 56 44 57 54 48 ...
## $ sex : int 1 1 0 1 1 0 1 1 1 0 ...
## $ cp : int 3 2 1 1 0 1 1 2 0 2 ...
## $ trtbps : int 145 130 130 120 140 140 120 150 140 130 ...
## $ chol : int 233 250 204 236 192 294 263 168 239 275 ...
## $ fbs : int 1 0 0 0 0 0 0 0 0 0 ...
## $ restecg : int 0 1 0 1 1 0 1 1 1 1 ...
## $ thalachh: int 150 187 172 178 148 153 173 174 160 139 ...
## $ exng : int 0 0 0 0 0 0 0 0 0 0 ...
```

```
## $ oldpeak : num 2.3 3.5 1.4 0.8 0.4 1.3 0 1.6 1.2 0.2 ...
## $ slp      : int 0 0 2 2 1 1 2 2 2 2 ...
## $ caa      : int 0 0 0 0 0 0 0 0 0 0 ...
## $ thall    : int 1 2 2 2 1 2 3 2 2 2 ...
## $ output   : int 1 1 1 1 1 1 1 1 1 1 ...
```

```
table(data$output, data$sex)
```

```
##
##      0  1
## 0 17 106
## 1 66  90
```

```
summary(table(data$output, data$sex))
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 26.704, df = 1, p-value = 2.371e-07
```

```
table(data$fbs, data$sex)
```

```
##
##      0  1
## 0 74 165
## 1  9  31
```

```
summary(table(data$fbs, data$sex))
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.1741, df = 1, p-value = 0.2786
```

```
table(data$cp, data$sex)
```

```
##
##      0  1
## 0 29 98
## 1 18 31
## 2 32 49
## 3  4 18
```

```
summary(table(data$cp, data$sex))
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 9.148, df = 3, p-value = 0.02739
```

```
table(data$restecg, data$sex)
```

```
##
##      0  1
## 0 36 100
## 1 45  96
## 2  2  0
```

```
summary(table(data$restecg, data$sex))
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 5.739, df = 2, p-value = 0.05673
##  Chi-squared approximation may be incorrect
```

```
test <- fisher.test(table(data$restecg , data$sex))
test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$restecg, data$sex)
## p-value = 0.06927
## alternative hypothesis: two.sided
```

```
table(data$slp , data$sex)
```

```
##
##      0  1
##  0  3 13
##  1 38 90
##  2 42 93
```

```
summary(table(data$slp, data$sex))
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1.0463, df = 2, p-value = 0.5927
##  Chi-squared approximation may be incorrect
```

```
test <- fisher.test(table(data$slp , data$sex))
test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(data$slp, data$sex)
## p-value = 0.6523
## alternative hypothesis: two.sided
```

```
table(data$thall , data$sex)
```

```
##
##      0  1
##  0  1  1
##  1  1 16
##  2 72 85
##  3  9 94
```

```
summary(table(data$thall, data$sex))
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 46.28, df = 3, p-value = 4.939e-10
##  Chi-squared approximation may be incorrect
```

```
test <- fisher.test(table(data$thall , data$sex))
test
```

```
##
## Fisher's Exact Test for Count Data
##
## data: table(data$thall, data$sex)
## p-value = 1.143e-11
## alternative hypothesis: two.sided
```

```
table(data$exng , data$sex)
```

```
##
##      0  1
## 0 68 124
## 1 15  72
```

```
summary(table(data$exng, data$sex))
```

```
## Number of cases in table: 279
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 9.464, df = 1, p-value = 0.002096
```

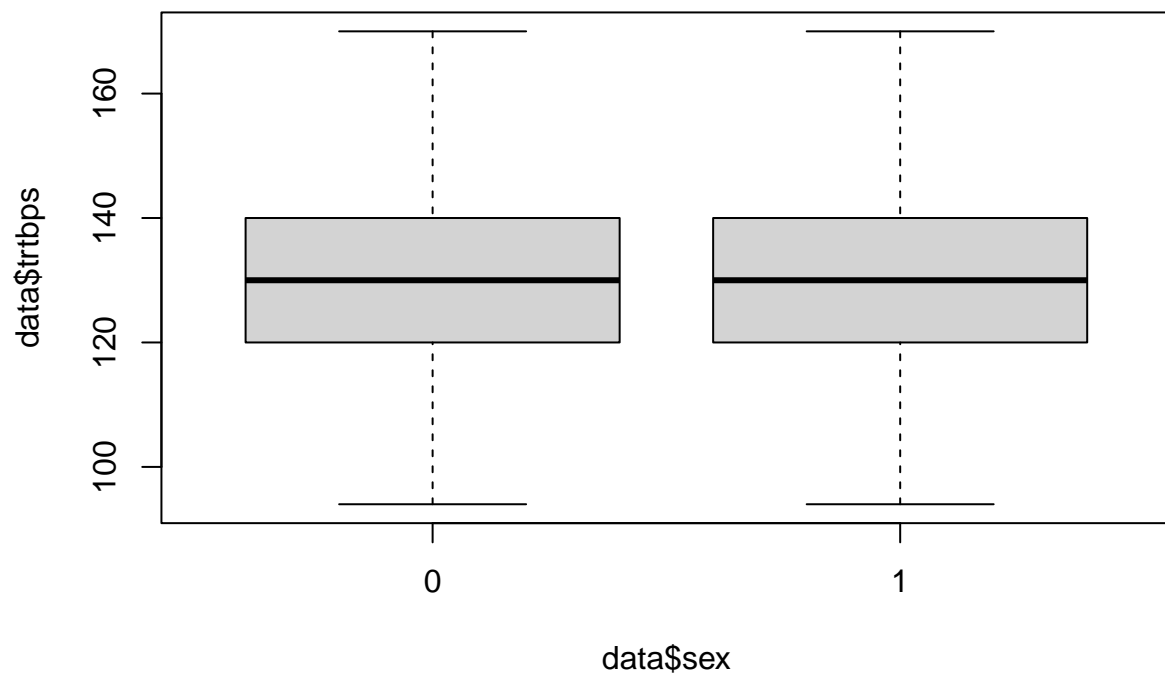
```
aggregate(chol~sex, data = data, FUN = var)
```

```
##   sex    chol
## 1   0 2116.052
## 2   1 1756.893
```

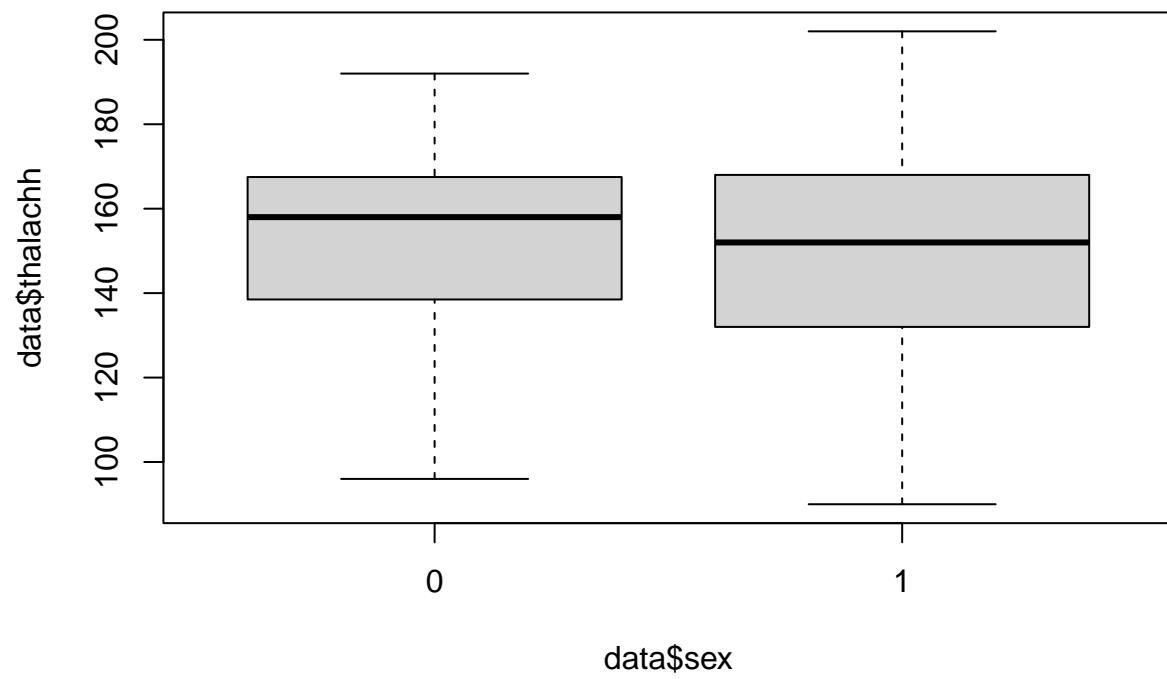
```
var.test(x = data[data$sex == 0, "chol"],
         y = data[data$sex == 1, "chol"] )
```

```
##
## F test to compare two variances
##
## data: data[data$sex == 0, "chol"] and data[data$sex == 1, "chol"]
## F = 1.2044, num df = 82, denom df = 195, p-value = 0.3008
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8462295 1.7624922
## sample estimates:
## ratio of variances
##      1.204429
```

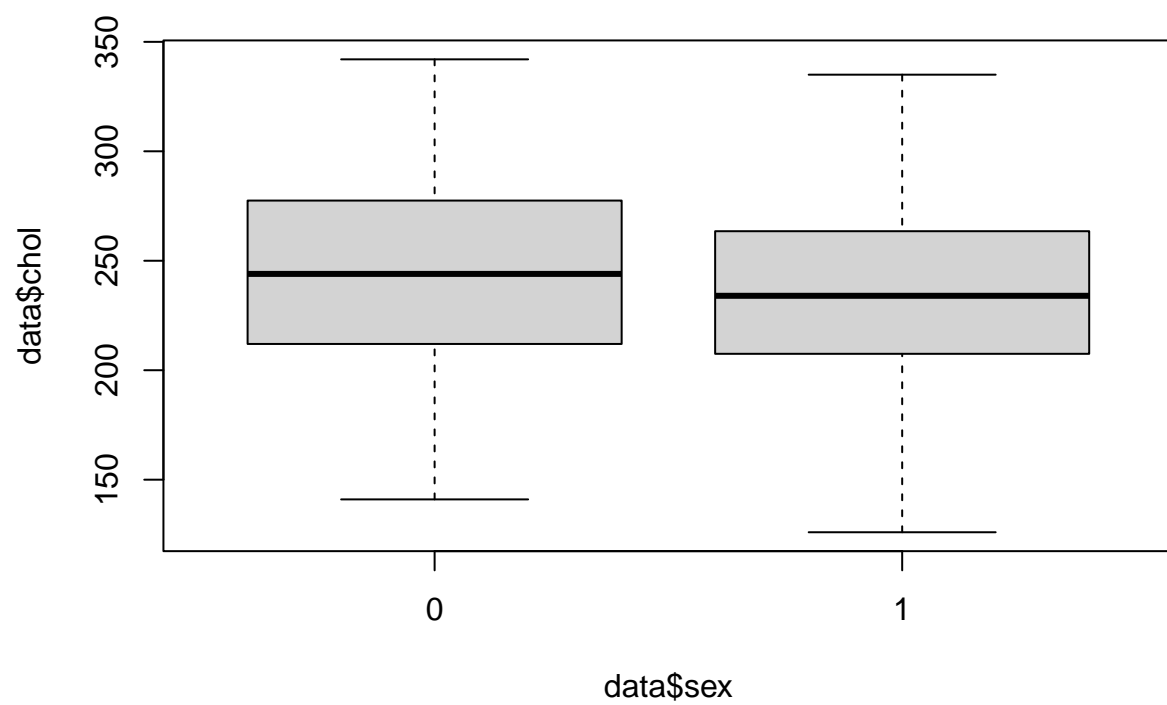
```
trtbps <- boxplot(data$trtbps ~ data$sex)
```



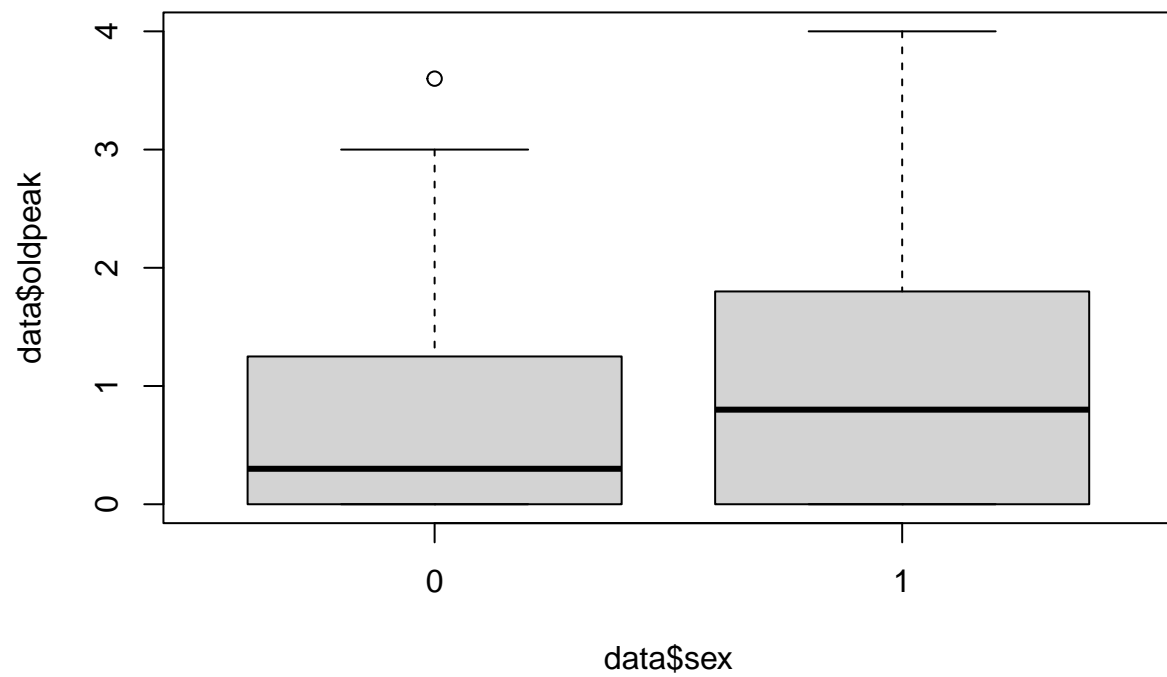
```
thalachh <- boxplot(data$thalachh ~ data$sex)
```



```
chol <- boxplot(data$chol ~ data$sex)
```

```
oldpeak <- boxplot(data$oldpeak ~ data$sex)
```



```
corr.res<-cor(data)
corrplot(corr.res,method="circle")
```

