PRACTICA 2 - Tipologia i cicle de vida de les dades

Autor: Lidia Toda i Sergi Garcia

Desembre 2022

Contents

Carreguem llibreries.

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

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 obtindre informació de quines son les variables més influents alhora de desenvolupar un atac de cor. De la mateixa manera, intentarem realitzar predicions sobre noves incorporacions de dades.

Llegim el dataset

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

Imprimim les primeres linies 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

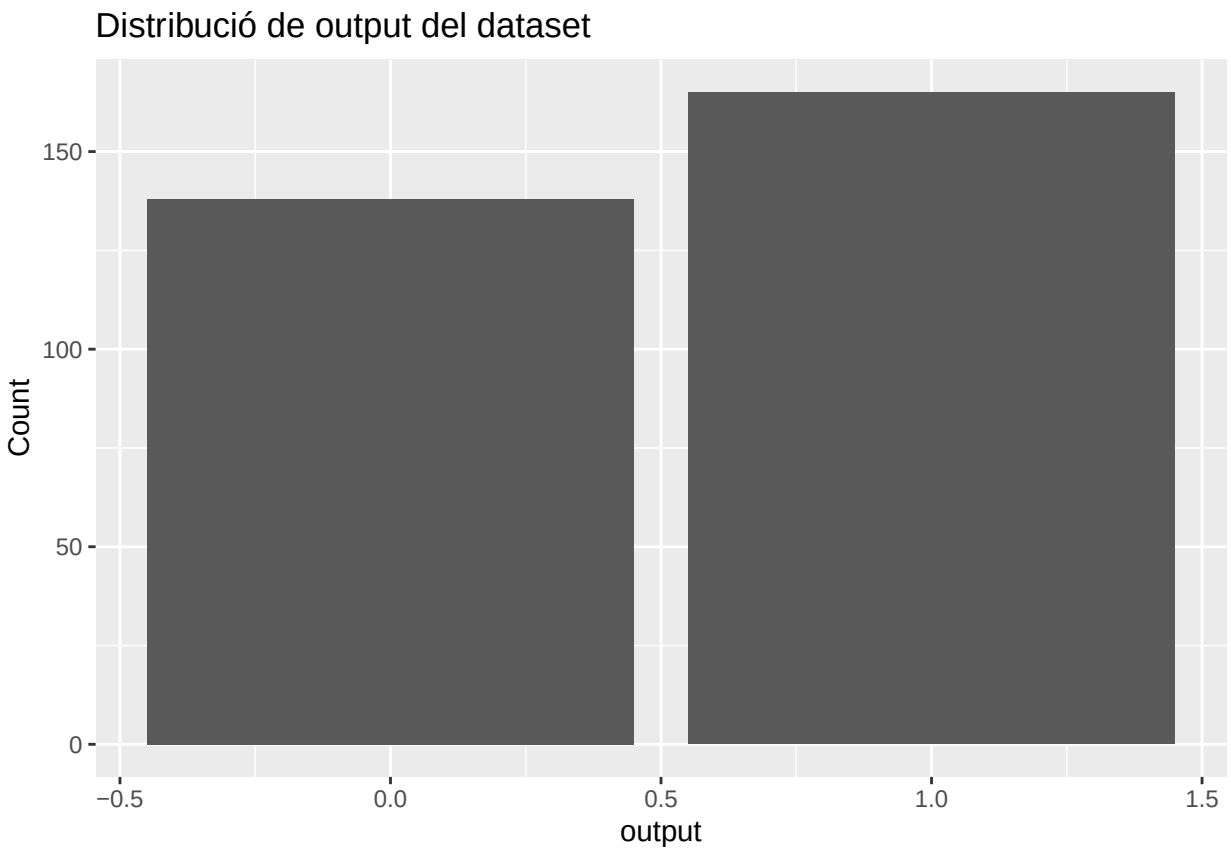
```
variables <- sapply(data,class)
kable(data.frame(variables=names(variables),clase=as.vector(variables)))
```

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
```

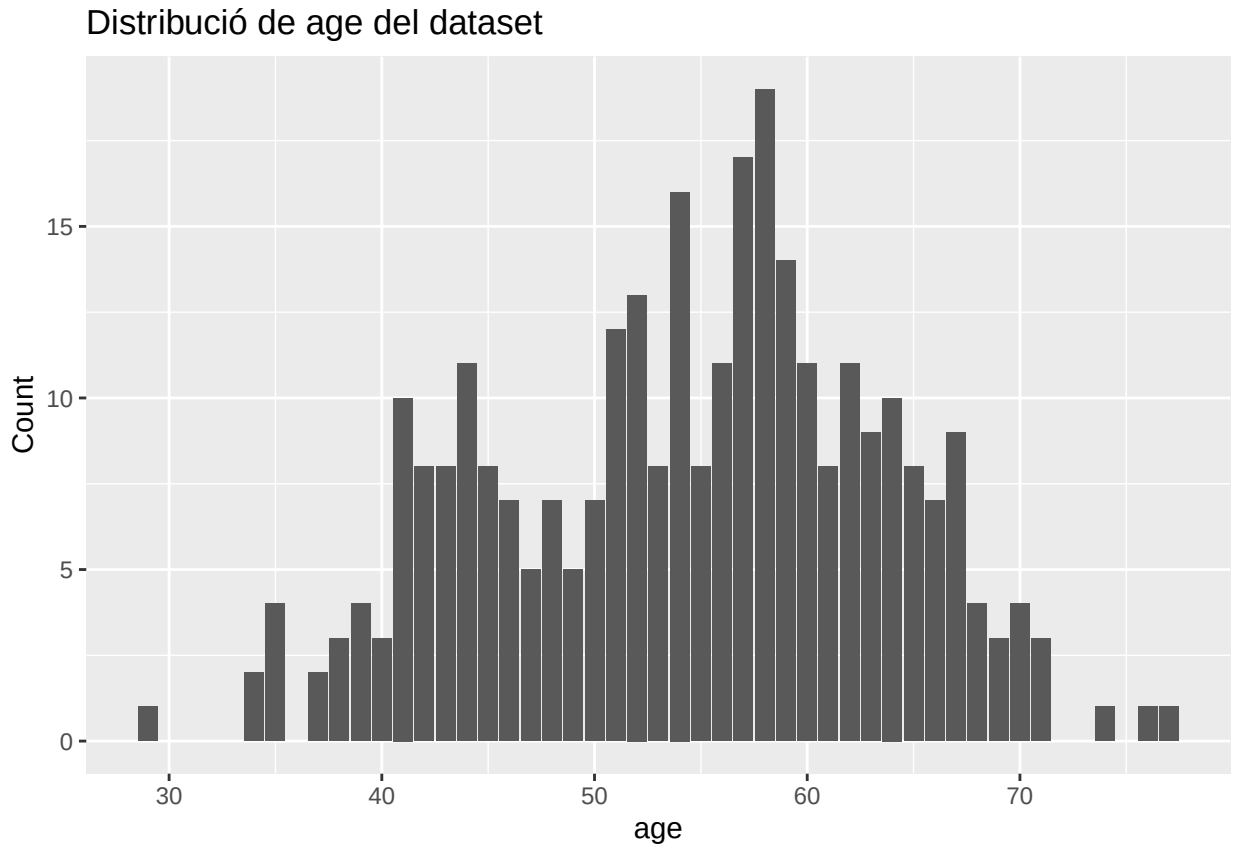


```r
output0 <- sum(data$output == 0)
output1 <- sum(data$output == 1)
percentatge_output0 = (output0/nombre_rows)
percentatge_output1 = (output1/nombre_rows)
cat("El percentatge de pacients amb output positiu és de", percentatge_output1, "mentre que el percentat
```

```
## El percentatge de pacients amb output positiu és de 0.5445545 mentre que el percentatge de pacients a
```

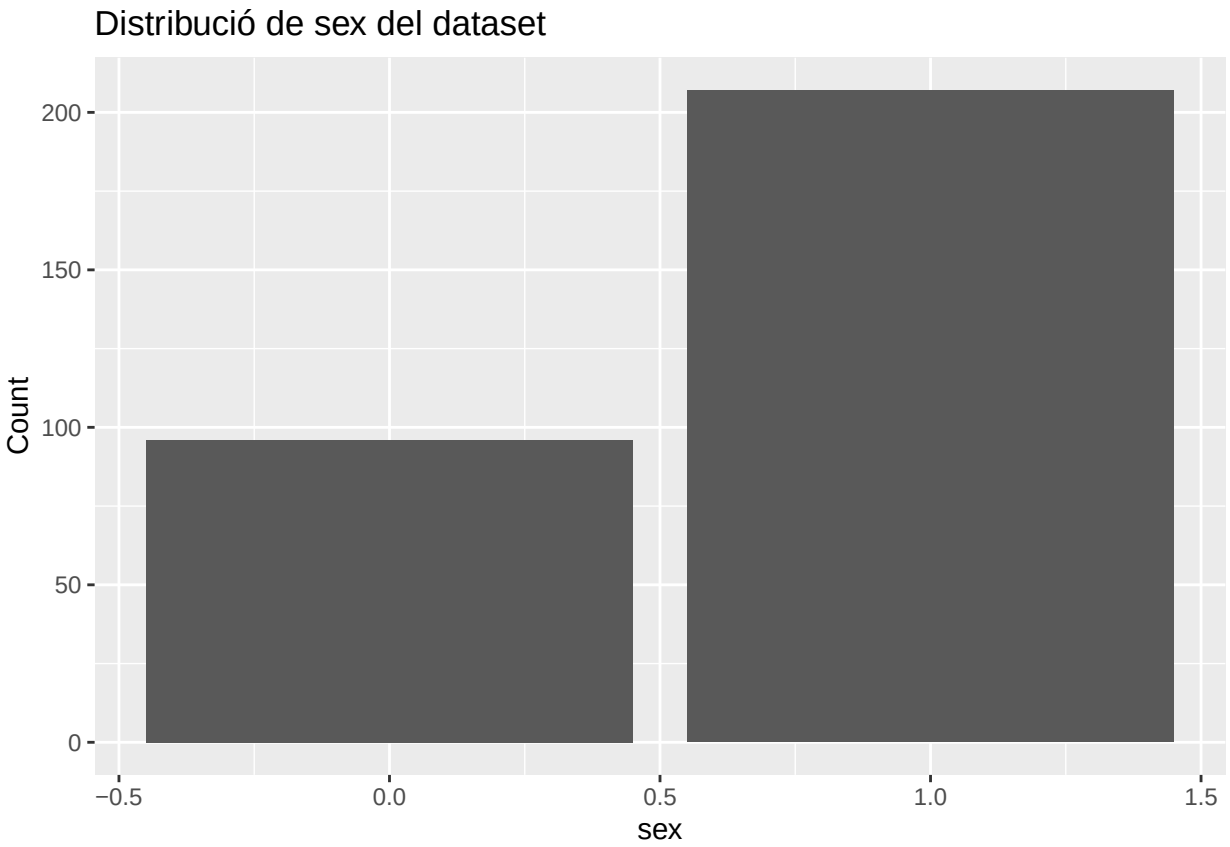
Observem la distribució de la columna age

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



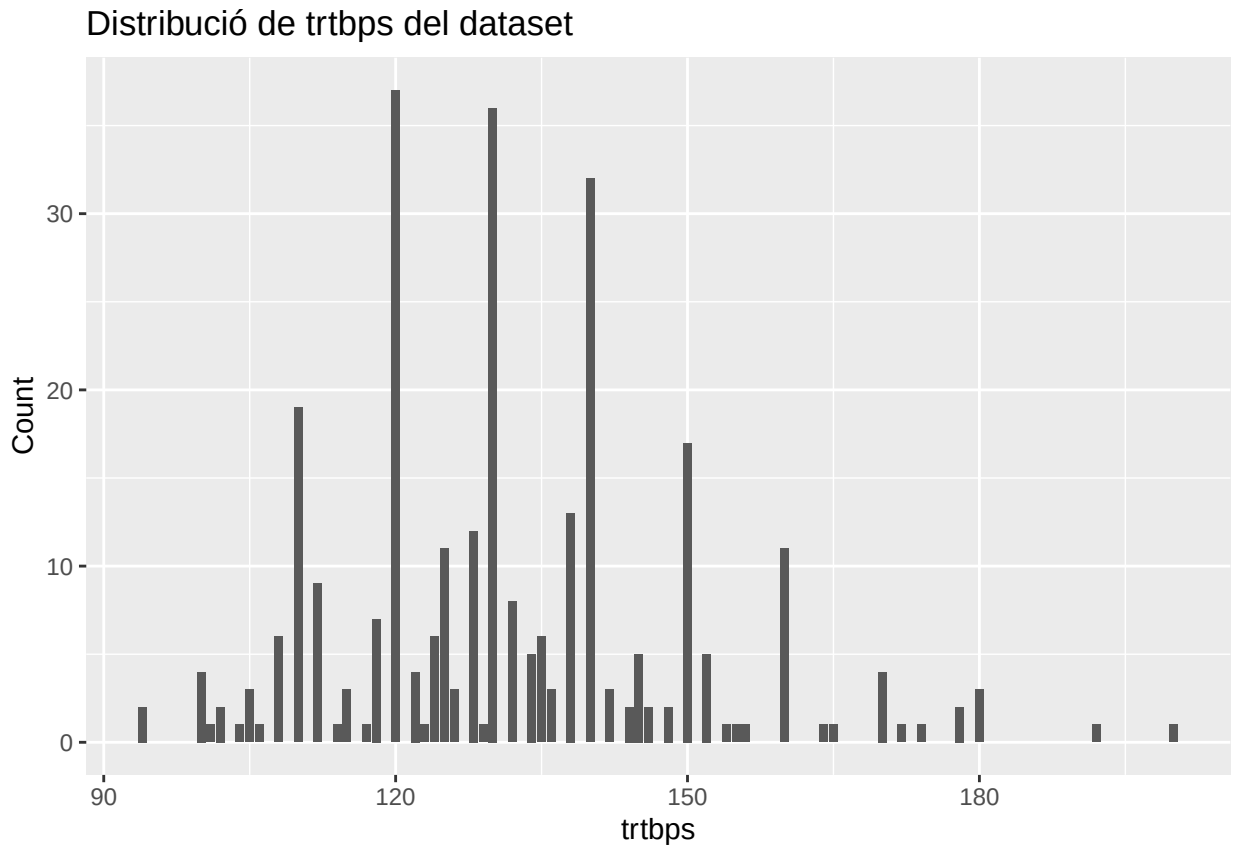
Observem la distribució de la columna sex

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



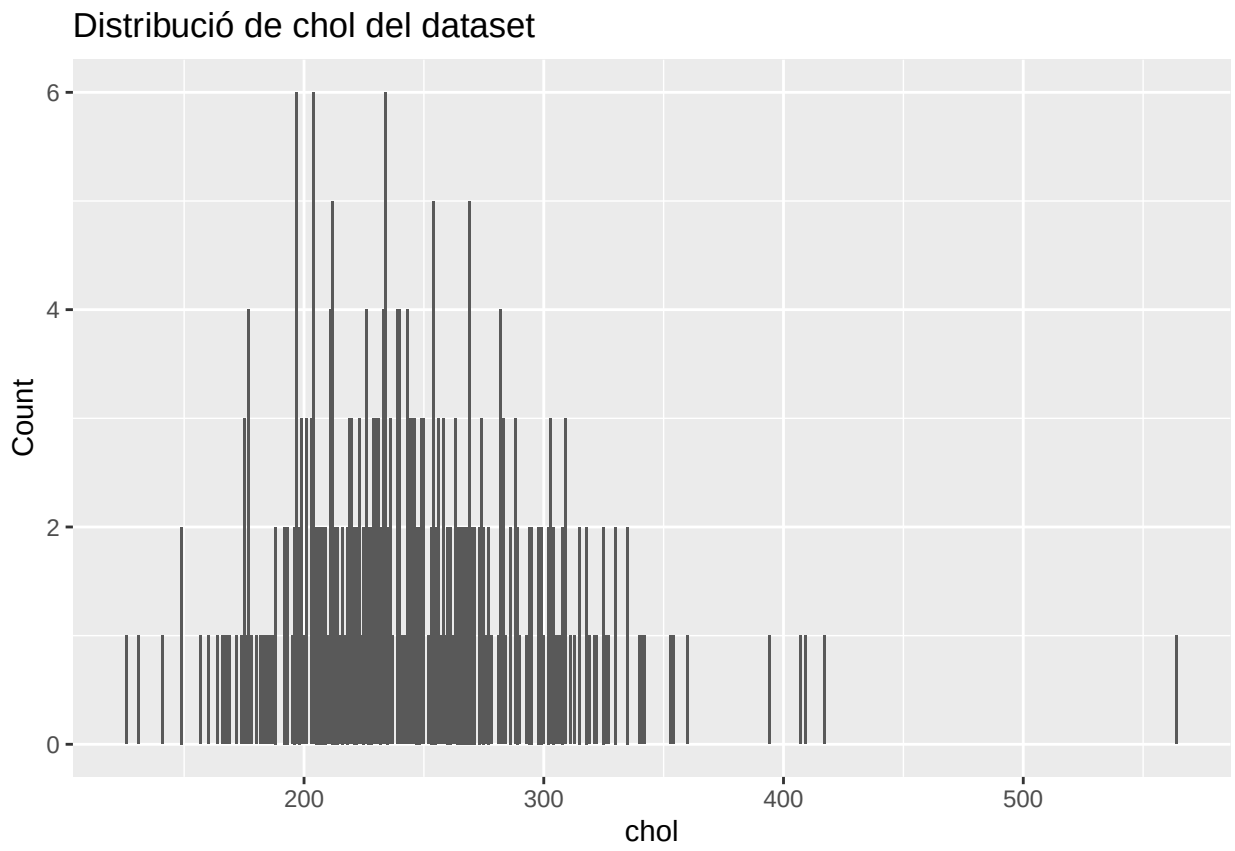
Observem la distribució de la columna trtbps

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



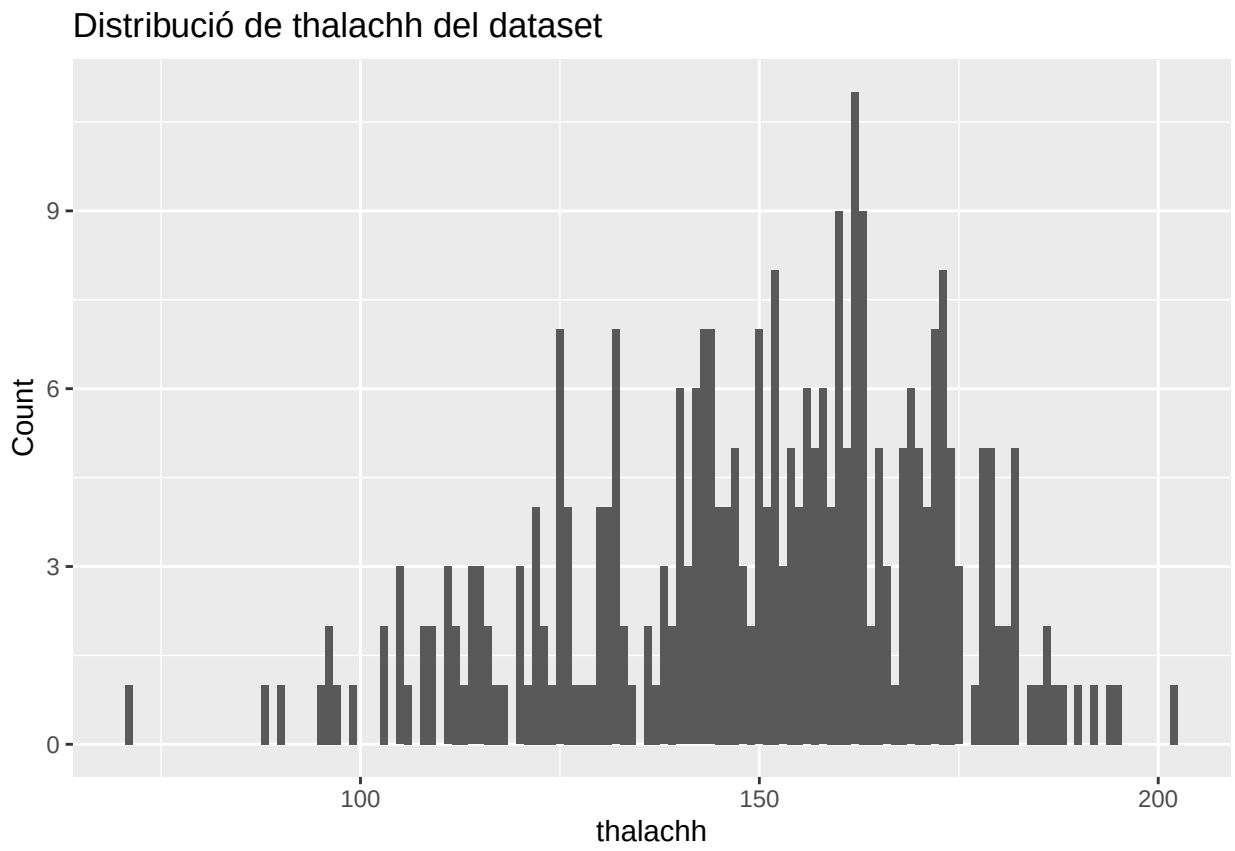
Observem la distribució de la columna chol

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



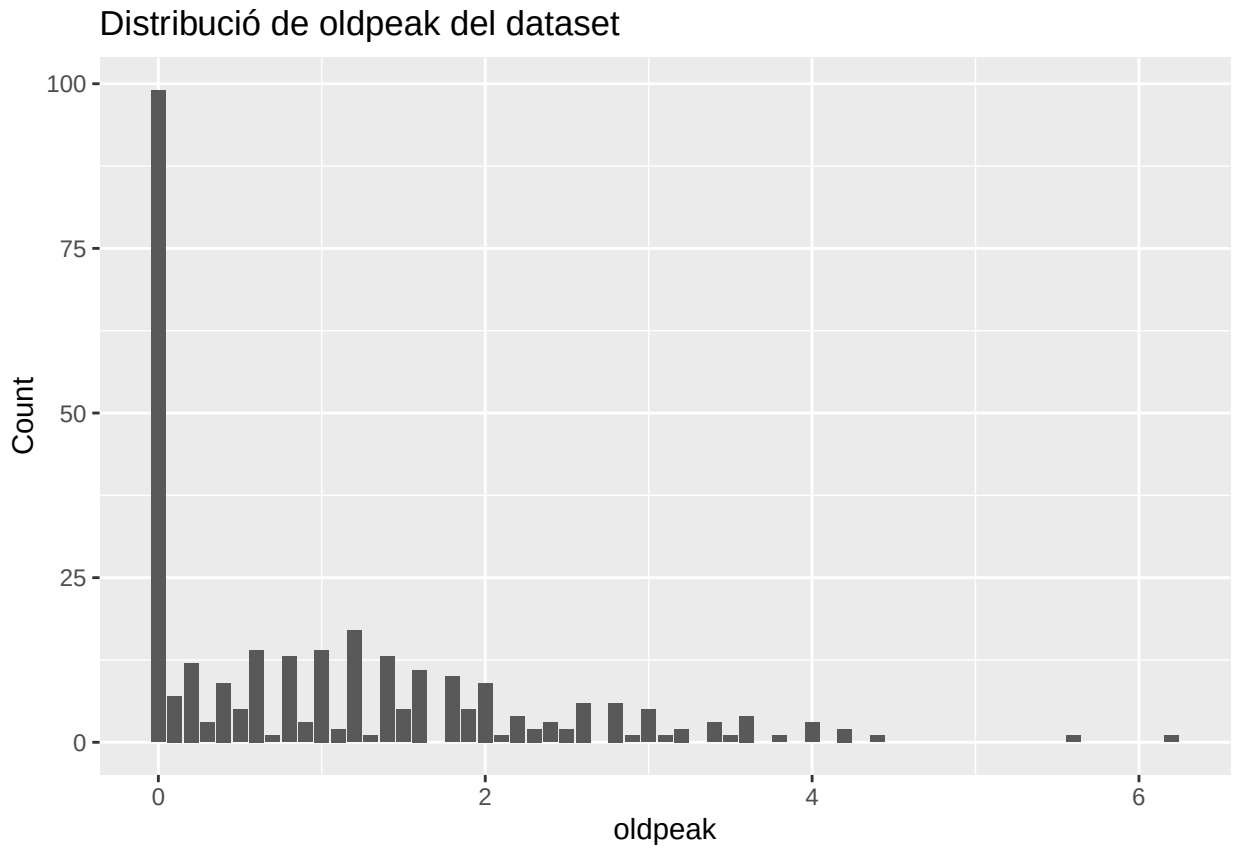
Observem la distribució de la columna thalachh

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



Observem la distribució de la columna oldpeak

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



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)
```

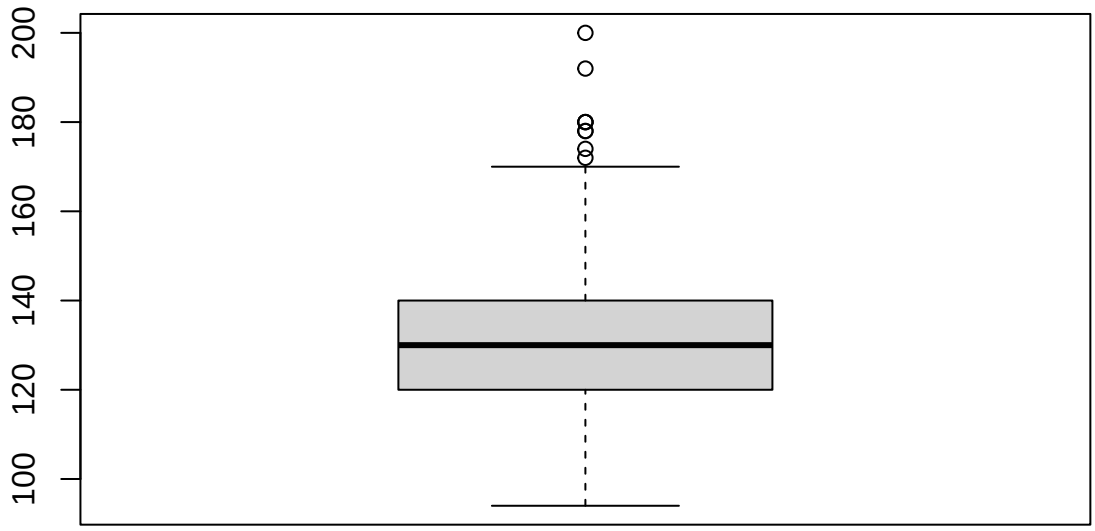
Obtenim informació de les noves rows i columns.

```
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)
```

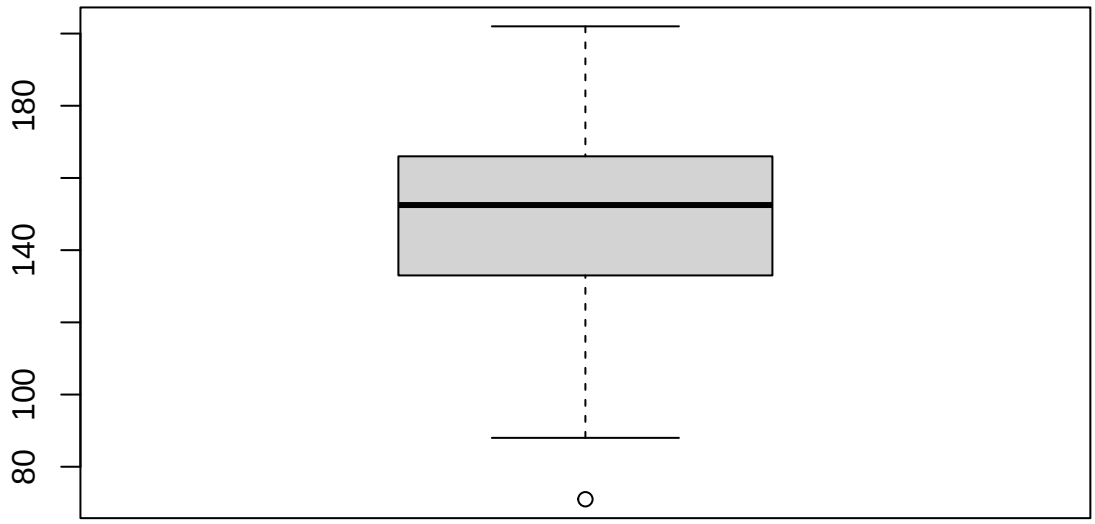
```
## El nombre de files és de 302 i el nombre de columnes és de 14
```

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

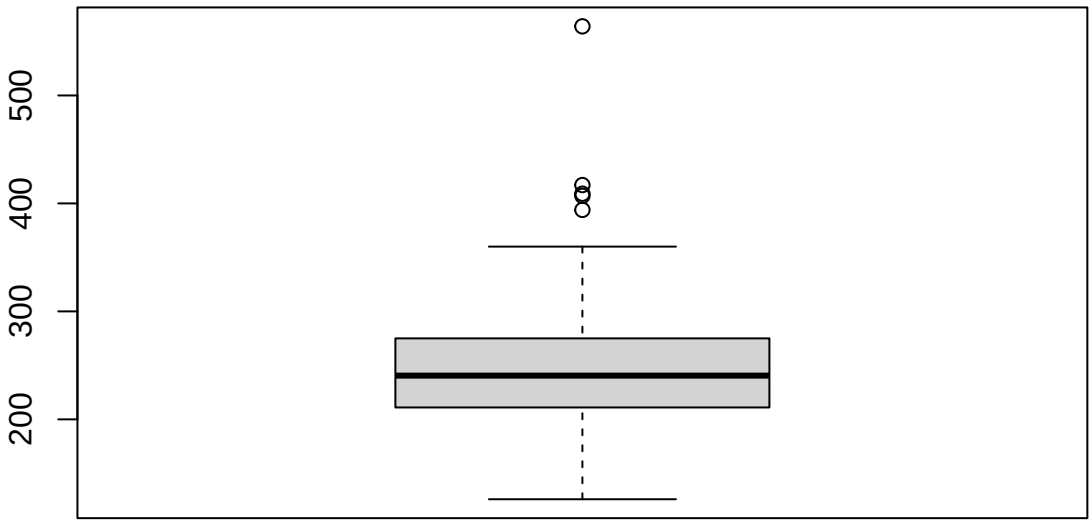
```r
trtbps <- boxplot(data$trtbps)
```



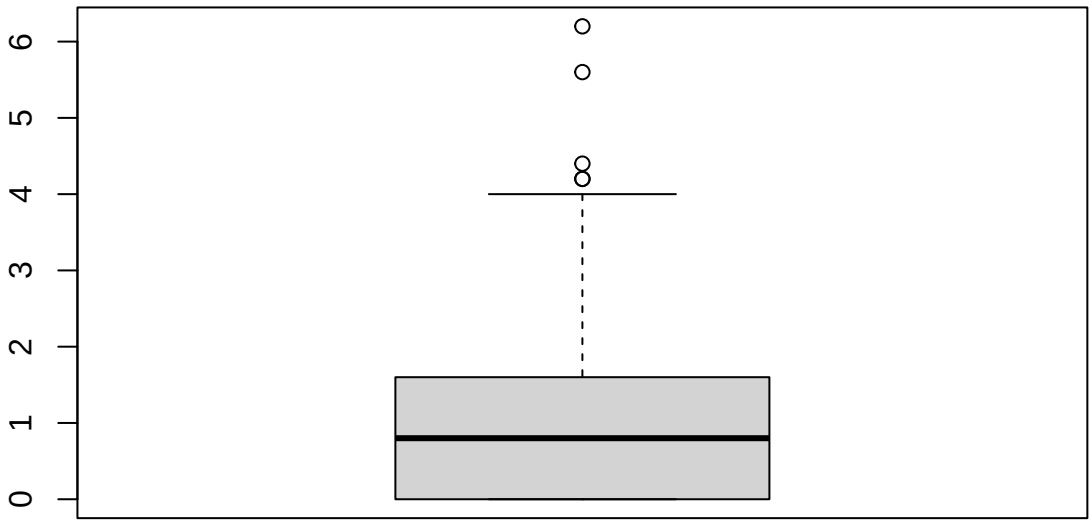

```r
thalachh <- boxplot(data$thalachh)
```




```
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)
```

Imprimim una taula de correlacions

```r
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

```r
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
```

```r
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("0.01067", "0.003466", "0.1485", "6.43e-05", "1.54e-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	0.01067	0.003466	0.1485	6.43e-05	1.54e-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
```

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(model, 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", "Si")
  )
colnames(mc_sl)<-c("Real","Predicció","Diferencia?")
kable(mc_sl)
```

	Real	Predicció	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

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