Heart Attack

Jaime Ulayar Arroyo

2023-02-08

Loading data

```
library(readr)
data <- read_delim("Data/heart_mod_2023-02-08(1).csv", delim = "p", escape_double = FALSE,locale = loca
```

```
## New names:
## Rows: 303 Columns: 15
## -- Column specification
## -------------------------------------------------------- Delimiter: "p" chr
## (1): target dbl (14): ...1, age, sex, cp, trestbps, chol, fbs, restecg,
## thalach, exang, ...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
head(data)
```

```
## # A tibble: 6 x 15
##    ...1   age   sex    cp trestbps  chol   fbs restecg thalach exang oldpeak
##   <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl>
## 1     1    63     1     3      145   233     1       0     150     0     2.3
## 2     2    37     1     2      130   250     0       1     187     0     3.5
## 3     3    41     0     1      130   204     0       0     172     0     1.4
## 4     4    56     1     1      120   236     0       1     178     0     0.8
## 5     5    57     0     0      120   354     0       1     163     1     0.6
## 6     6    57     1     0      140   192     0       1     148     0     0.4
## # ... with 4 more variables: slope <dbl>, ca <dbl>, thal <dbl>, target <chr>
```

En el campo de edad aparecen 3 registros con edades imposibles, seguramente debidas a la presencia de una coma decimal, ya que acaban las 3 en 0.

En el campo resting bloodpressure hay un valor anormalmente alto. De nuevo parece ser por la presencia de una coma decimal.

En el campo de colesterol hay un valor particularmente bajo.

En el campo target hay un valor "11" y un valor "O", cuando es un campo lógico. Lo más probable es que correspondieran a los valores "1" y "0" respectivamente.

Renaming

```
data$sex <-  as.factor(data$sex)
levels(data$sex) <- c("Female", "Male")

data$cp <- as.factor(data$cp)
levels(data$cp) <- c("Assymmptomatic", "Atypical angina", "No angina", "Typical angina")

data$fbs <- as.factor(data$fbs)
levels(data$fbs) <- c("No", " Yes")

data$restecg <- as.factor(data$restecg)
levels(data$restecg) <- c("Hypertrophy", "Normal", "Abnormalities")

data$exang <- as.factor(data$exang)
levels(data$exang) <- c("No", "Yes")

data$slope <- as.factor(data$slope)
levels(data$slope) <- c("Descending", "Flat", "Ascending")

data$thal <- as.factor(data$thal)
levels(data$thal) <- c("Fixed defect","Normal flow", "Reversible defect")

data$target <- as.character(data$target)

data$target[data$target == "11"] <- 1
data$target[data$target == "O"] <- 0

data$target <- as.factor(data$target)
levels(data$target) <- c("Yes", "No")

head(data)
```

```
## # A tibble: 6 x 15
##    ...1   age sex    cp        trest~1  chol fbs   restecg thalach exang oldpeak
##   <dbl> <dbl> <fct>  <fct>       <dbl> <dbl> <fct> <fct>     <dbl> <fct>   <dbl>
## 1     1    63 Male   Typical ~     145   233 " Ye~ Hypert~     150 No        2.3
## 2     2    37 Male   No angina     130   250 "No"  Normal      187 No        3.5
## 3     3    41 Female Atypical~     130   204 "No"  Hypert~     172 No        1.4
## 4     4    56 Male   Atypical~     120   236 "No"  Normal      178 No        0.8
## 5     5    57 Female Assymmpt~     120   354 "No"  Normal      163 Yes       0.6
## 6     6    57 Male   Assymmpt~     140   192 "No"  Normal      148 No        0.4
## # ... with 4 more variables: slope <fct>, ca <dbl>, thal <fct>, target <fct>,
## #   and abbreviated variable name 1: trestbps
```

Separating data

```
data <- data[,-1]

type_class <- sapply(data, class)
table(type_class)
```

```
## type_class
##  factor numeric
##       8       6
```

```
data$oldpeak <- gsub(",", ".", data$oldpeak)
data$oldpeak <- as.numeric(data$oldpeak)

data_num <- data[,type_class %in% c("integer", "numeric")]
data_fac <- data[,type_class %in% c("factor")]
```

Checking

```
data_num$chol[data_num$chol == 5] <- NA
data_num$trestbps[data_num$trestbps == 1540] <- 154
data_num$age[data_num$age >= 100] <- data_num$age[data_num$age >= 100]/10

meantrestbps <- mean(data$trestbps,na.rm = TRUE)
data_num$trestbps[is.na(data_num$trestbps)] <- meantrestbps

meanchol <- mean(data$chol,na.rm = TRUE)
data_num$chol[is.na(data_num$chol)] <- meanchol

meanthalach <- mean(data$thalach,na.rm = TRUE)
data_num$thalach[is.na(data_num$thalach)] <- meanthalach

meanca <- mean(data$ca,na.rm = TRUE)
data_num$ca[is.na(data_num$ca)] <- meanca

summary(data_num)
```

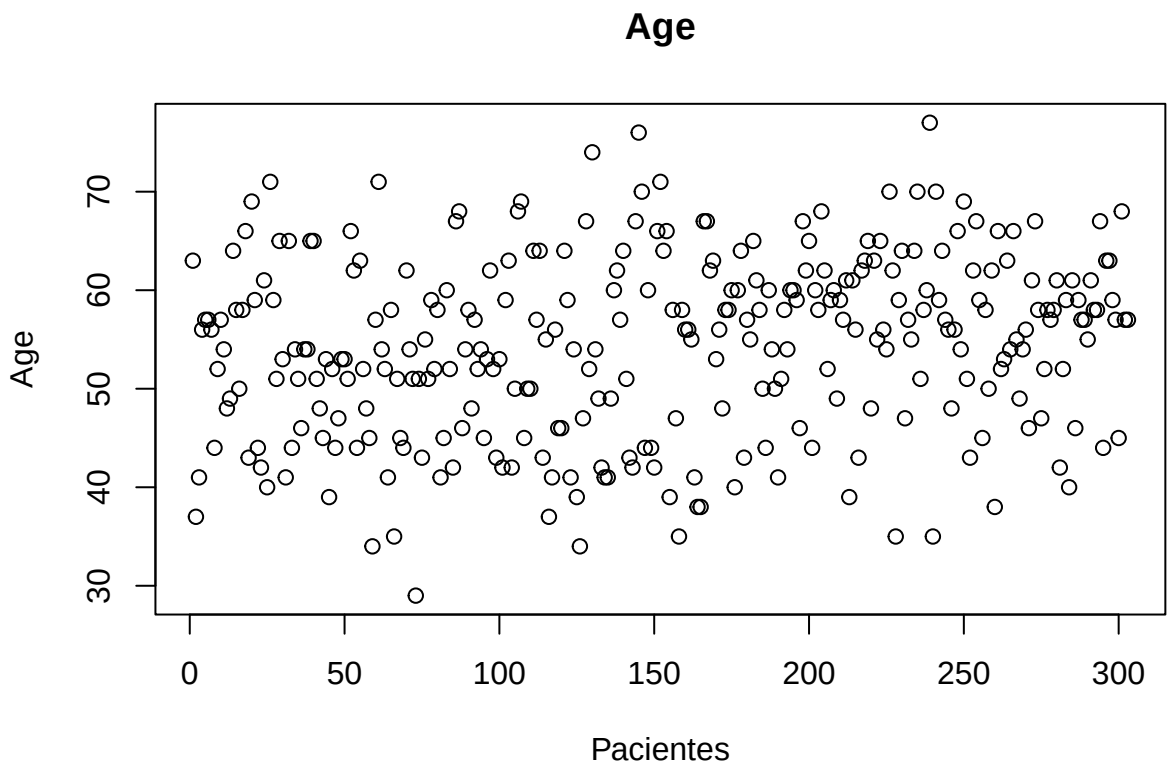
```
##       age           trestbps          chol          thalach         oldpeak
##  Min.   :29.00   Min.   : 94.0   Min.   :126.0   Min.   : 71.0   Min.   :0.00
##  1st Qu.:47.50   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:133.5   1st Qu.:0.00
##  Median :55.00   Median :130.0   Median :240.0   Median :152.0   Median :0.80
##  Mean   :54.37   Mean   :131.6   Mean   :245.2   Mean   :149.6   Mean   :1.04
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:274.0   3rd Qu.:166.0   3rd Qu.:1.60
##  Max.   :77.00   Max.   :200.0   Max.   :417.0   Max.   :202.0   Max.   :6.20
##        ca
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.6745
##  3rd Qu.:1.0000
##  Max.   :3.0000
```

Visualization

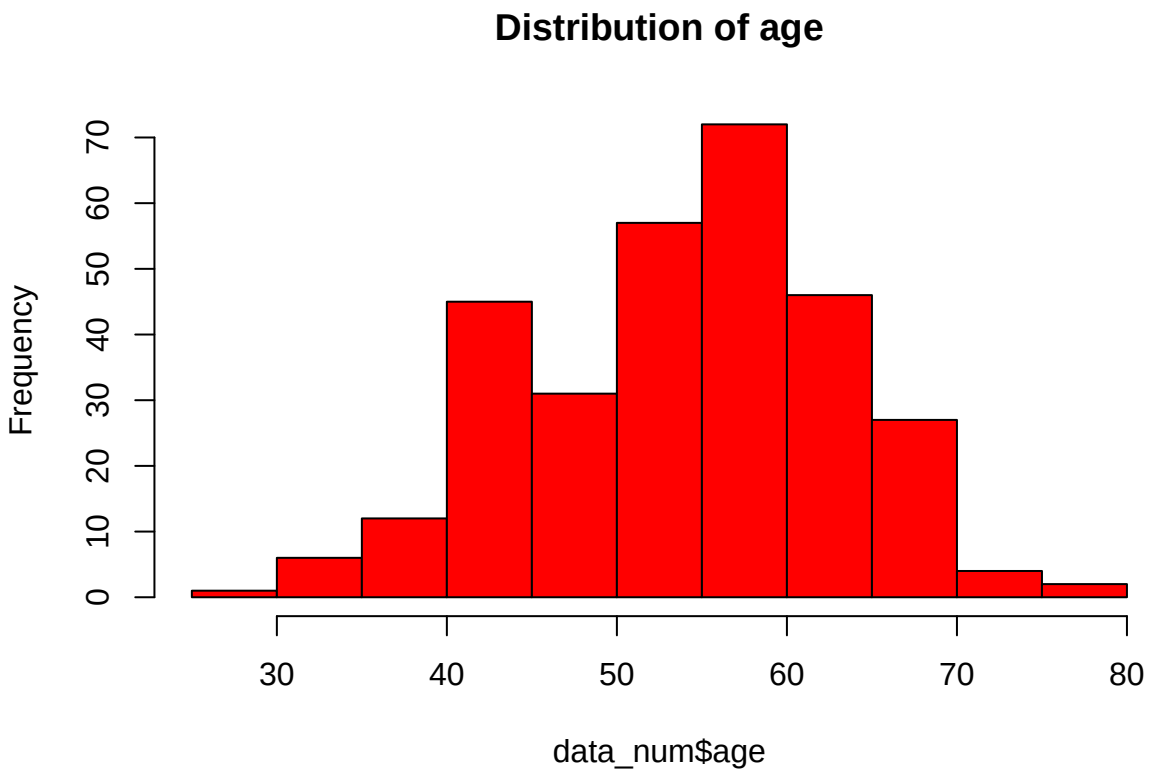
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v dplyr   1.0.10
## v tibble  3.1.8      v stringr 1.4.1
## v tidyr   1.2.1      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

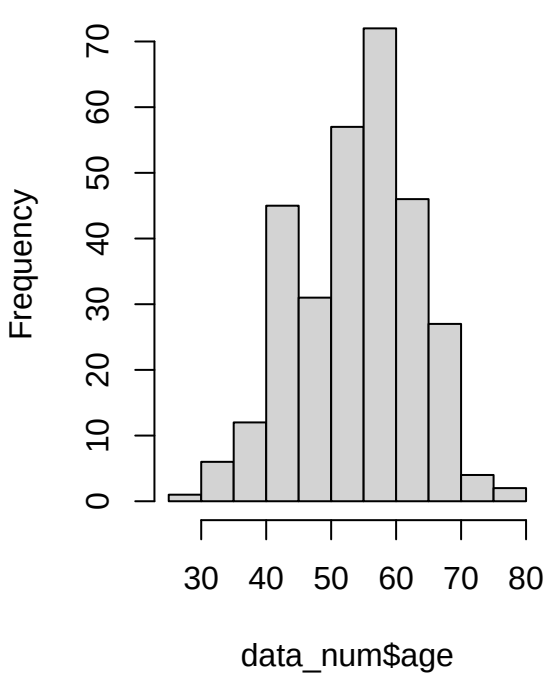
```r
plot(data_num$age, main = "Age", xlab = "Pacientes", ylab = "Age")
```

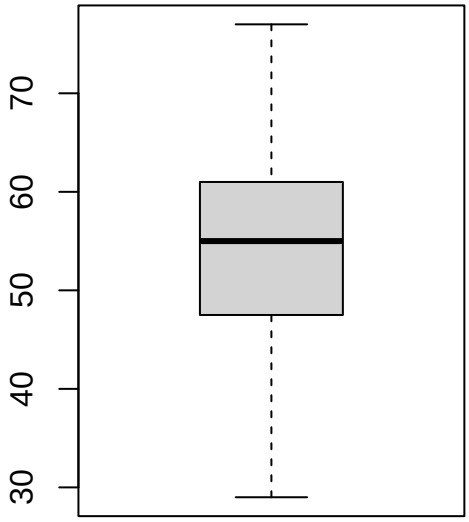



```r
hist(data_num$age, col = "red", breaks = (max(data_num$age)-min(data_num$age))/5, main = "Distribution
```




```
par(mfrow = c(1,2))
hist(data_num$age)
boxplot(data_num$age)
```

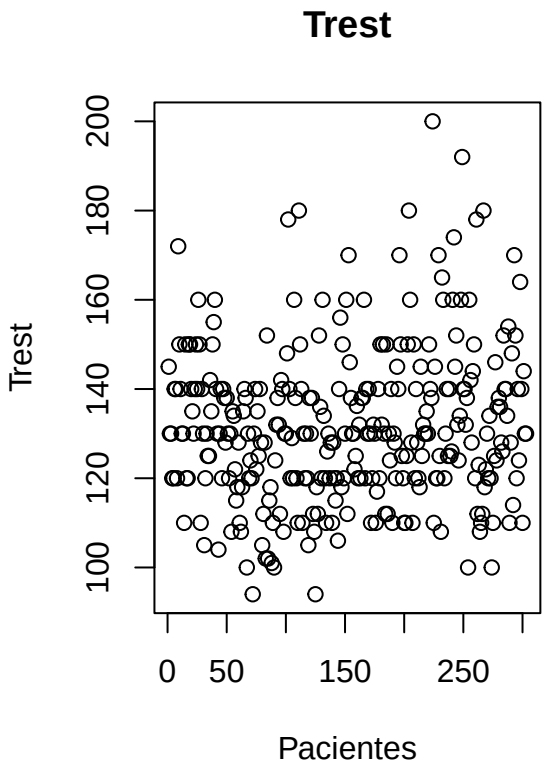


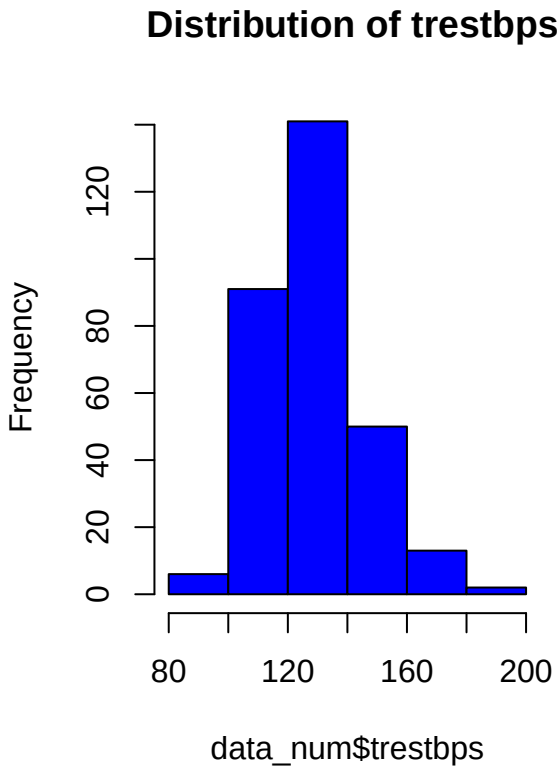



```
shapiro.test(data_num$age)
```

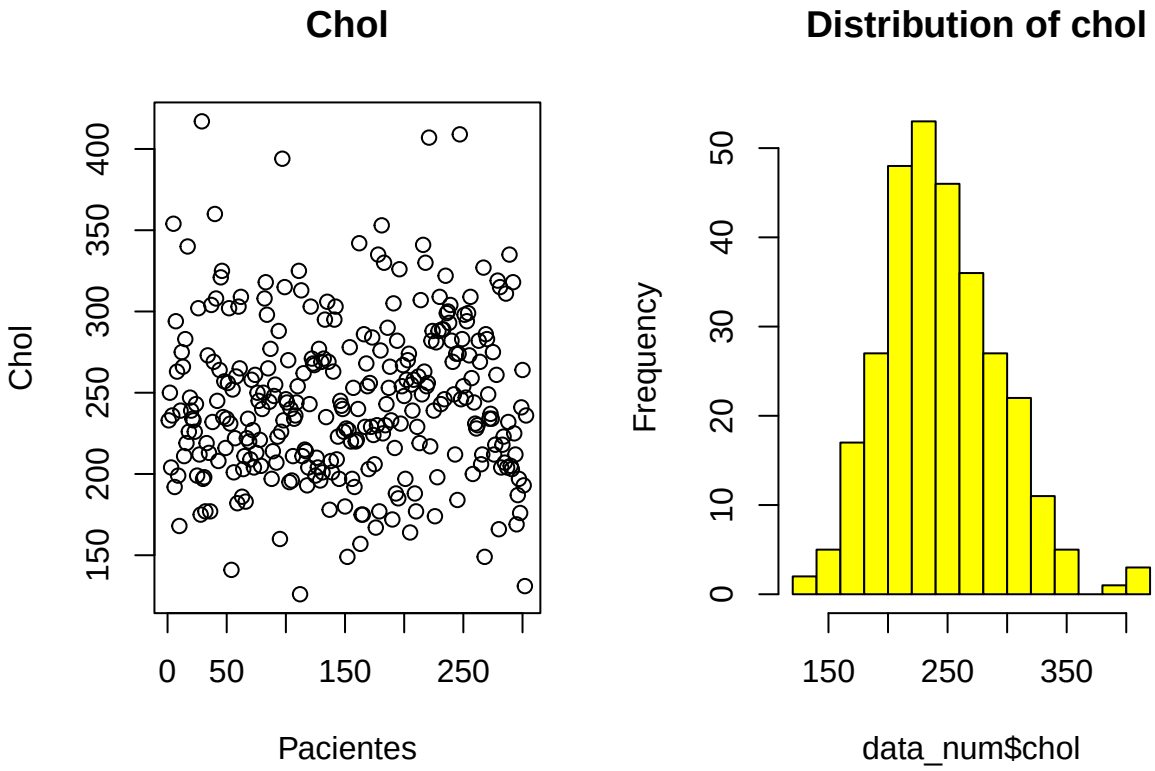
```
##
##  Shapiro-Wilk normality test
##
## data:  data_num$age
## W = 0.98637, p-value = 0.005798
```

```
plot(data_num$trestbps, main = "Trest", xlab = "Pacientes", ylab = "Trest")
hist(data_num$trestbps, col = "blue", breaks = (max(data_num$trestbps)-min(data_num$trestbps))/20, main
```

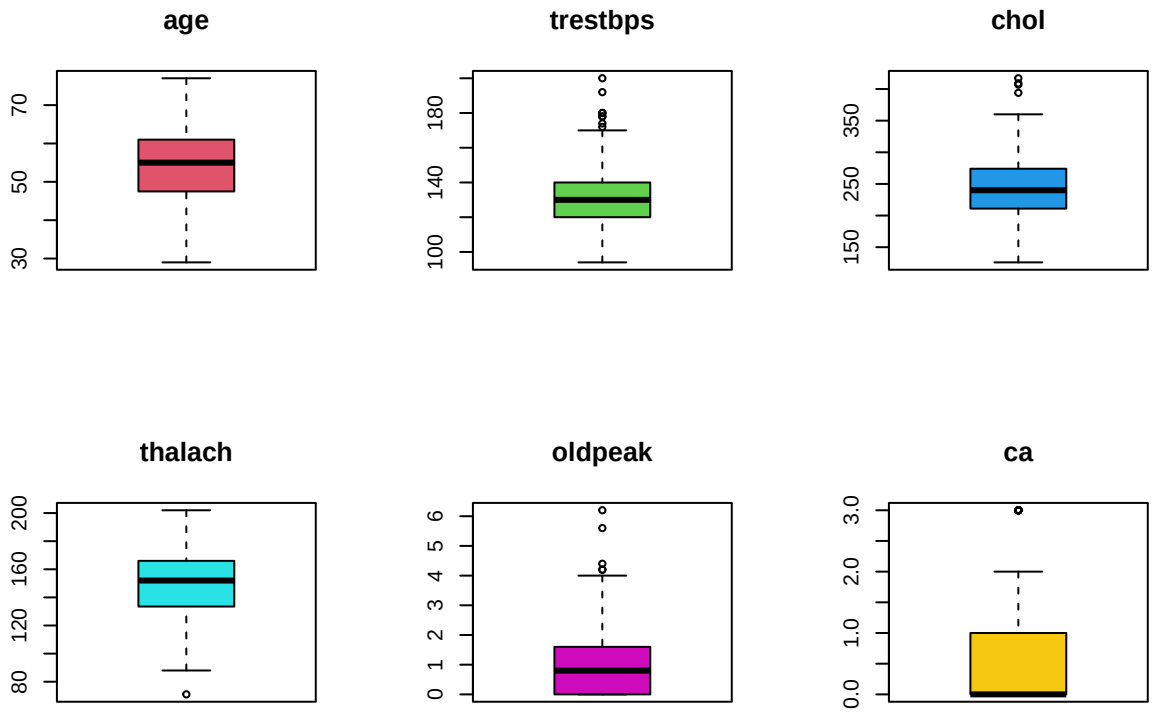




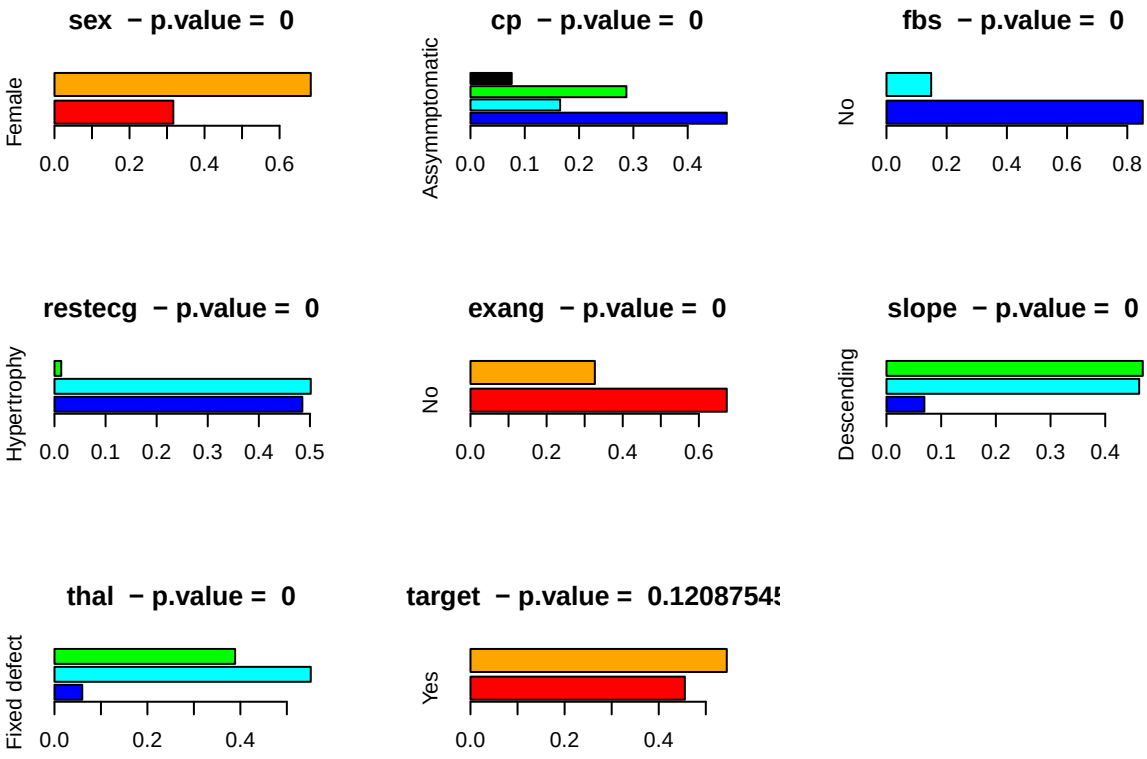

```
plot(data_num$chol, main = "Chol", xlab = "Pacientes", ylab = "Chol")
hist(data_num$chol, col = "yellow", breaks = (max(data_num$chol)-min(data_num$chol))/20, main = " Distri
```




```
par(mfrow=c(2,3))
for (i in 1:length(data_num)){
  boxplot(data_num[i], main = names(data_num)[i], col = i+1)
}
```




```
par(mfrow = c(3,3))
for (i in 1:length(data_fac)){
  testchis = chisq.test(table(data_fac[i]))
  if (testchis$p.value > 1e-15){
    color = c("red", "orange", "yellow", "white")
  } else {
    color = c("blue", "cyan", "green", "black")
  }
  barplot(prop.table(table(data_fac[[i]])), main = paste(names(data_fac)[i]," - p.value = ", round(test
}
```


sex − p.value = 0
Female
0.0 0.2 0.4 0.6
cp − p.value = 0
Assymmptomatic
0.0 0.1 0.2 0.3 0.4
fbs − p.value = 0
No
0.0 0.2 0.4 0.6 0.8
restecg − p.value = 0
Hypertrophy
0.0 0.1 0.2 0.3 0.4 0.5
exang − p.value = 0
No
0.0 0.2 0.4 0.6
slope − p.value = 0
Descending
0.0 0.1 0.2 0.3 0.4
thal − p.value = 0
Fixed defect
0.0 0.2 0.4
target − p.value = 0.1208754
Yes
0.0 0.2 0.4