Master en Ciencias de Datos - Tipología y ciclo de vida de los datos. Practica 2 - Limpieza y análisis de datos

Andrea Marcos Vargas y Felipe E García González

3 de junio de 2023

Índice

1 Descripción del dataset

El *dataset* se denomina *Heart Attack Analisys & Prediction*, nos proporciona 14 variables que podrían estar relacionadas con los ataques al corazón, para ello se nos presentan los datos de 303 personas. Este *dataset* nos parece importante porque puede intenta utilizarse para prevenir o predecir una enfermedad, todo lo relacionado con la salud de las personas es importante. El problema que se pretende resolver es la prevención de los ataques al corazón.

El *dataset* está disponible en Kaggle https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

En esta práctica se intentará obtener un modelo a partir de los datos disponibles que nos permita predecir la probabilidad de que una persona sufra un ataque al corazón en base a los valores de ciertas variables disponibles. Se intentará que el modelo sea lo más sencillo posible, utilizando en el modelo aquellas variables que realmente sean significativas para el modelo y descartando aquellas que no aporten información al mismo.

Las variables que tenemos en el *dataset* son:

- **age**: Edad de la persona, variable cuantitativa discreta.
- **sex**: Sexo de la persona, variable cualitativa nominal. Toma valor 0 para mujer y 1 para hombre.
- **cp**: (Chest Pain), define el tipo de dolor en el pecho que presenta la persona, variable cualitativa o categórica ordinal con 4 categorías. Toma valor 0 para dolor típico, 1 para dolor atípico, 2 sin dolor y 3 asintomático.
- **trtbps**: Presión arterial en reposo (en mm Hg), variable cuantitativa discreta.
- **chol**: Colesterol en mg/dl, variable cuantitativa discreta.
- **fbs**: Azúcar en sangre mayor a 120 mg/dl, variable cualitativa dicotómica. Toma valor 1 para *True* y 0 para *False*.
- **restegc**: Tipo de electrocardiograma en descanso. Variable cualitativa o categórica ordinal con tres categorías. Toma valor 0 para normal, 1 para onda con anormalidad y para 2 aparición de hipertrofia en el ventrículo izquierdo.
- **thalach**: Valor máximo de pulsaciones cardíacas detectadas, variable cuantitativa discreta.
- **exng**: Angina inducida por el ejercicio, variable cualitativa dicotómica. Toma valor 1 para *si* y 0 para *no*.
- **oldpeak**: Valor de ST depresión, inducida por el ejercicio relativo al reposo, variable cuantitativa continua.
- **slp**: (slope) Pendiente del segmento ST, variable continua discreta.
- **caa**: Número de venas que se pueden ver en una fluoroscopia, variable continua discreta.
- **thall**: Nivel de talasemia, variable continua discreta.
- **output**: Si el paciente ha sufrido un ataque al corazón o no.

```
datos <- read.table("heart.csv", header= TRUE, sep = ",", dec = ".")
datos %>% str()
```

```
## '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 ...
```

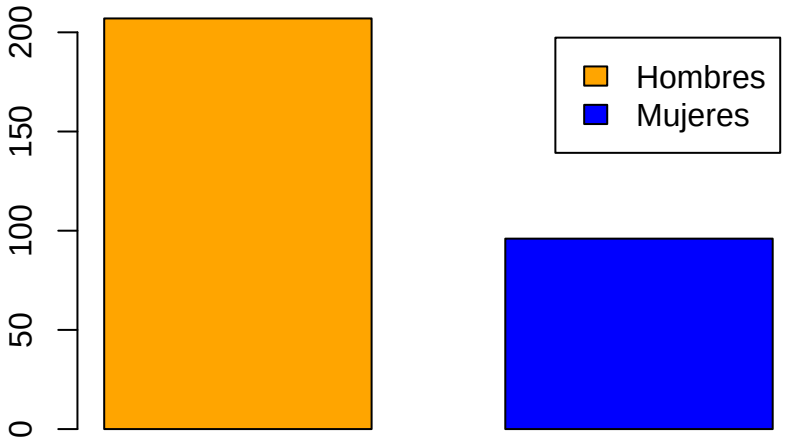
```
clases <- sapply(datos, class)
kable(data.frame(variables=names(clases), clase=as.vector(clases)))
```

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

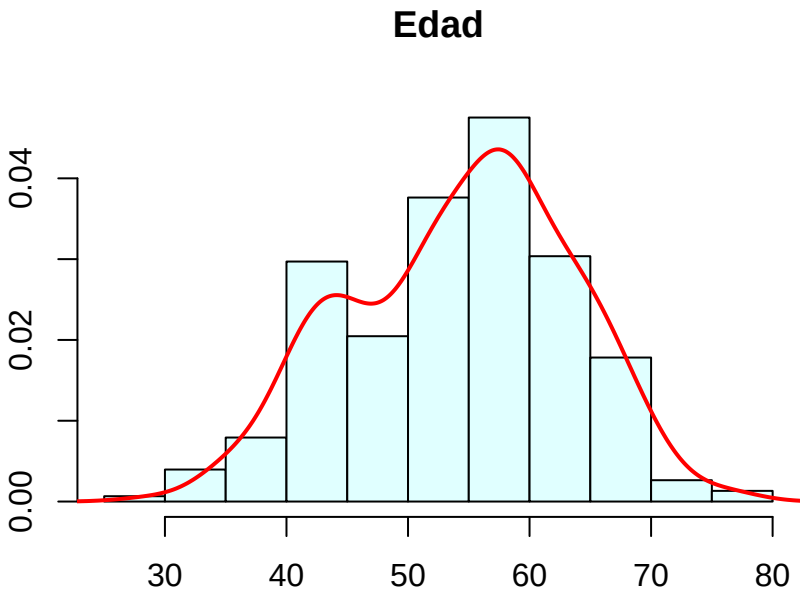
```
# Número de hombres y mujeres
h <- length(which(datos$sex==1)) ; m <- length(which(datos$sex==0))
```

El *dataset* tiene un tamaño de 303 filas y 14 columnas, el número de mujeres es de 96 y el de hombres 207, y la distribución por sexos y edad es:

```
barplot(c(h,m), col = c("orange", "blue"), legend.text = c("Hombres", "Mujeres"), space = c(0.5,0.5))
```



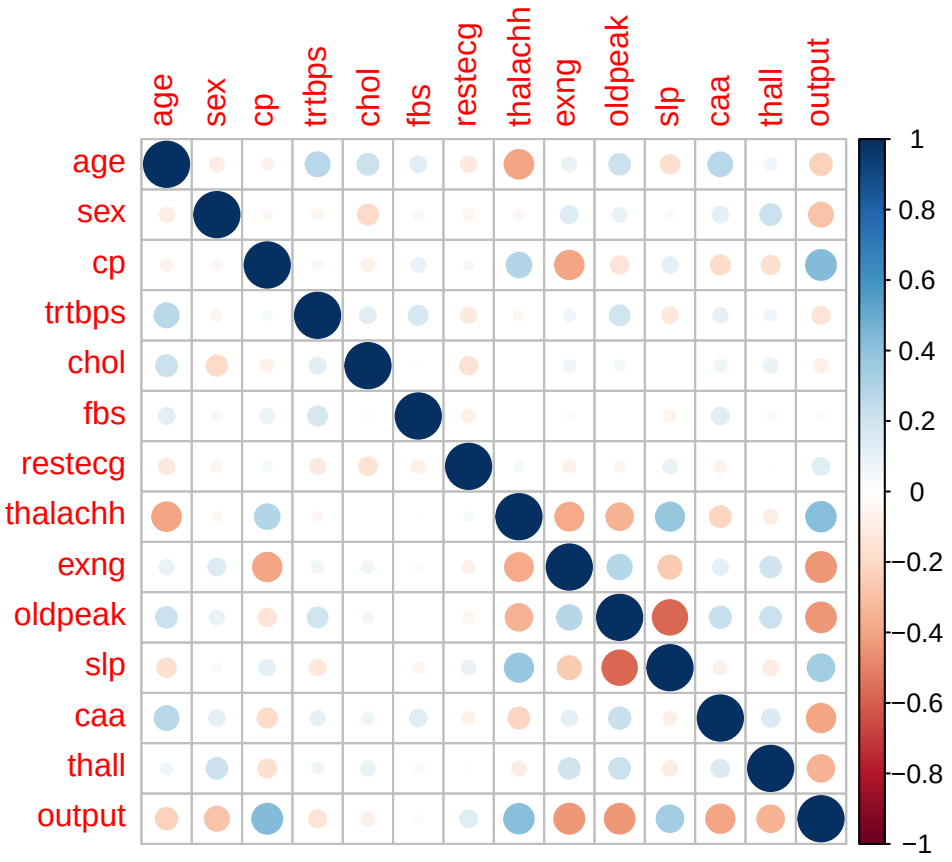

```
hist(datos$age,freq=FALSE, col="lightcyan", main="Edad",xlab="",ylab="")
lines(density(datos$age),col="red",lwd=2)
```



2 Integración y selección

En este apartado vamos a analizar la relaciones entre las distintas variables del *dataset* con *output*, que es la variable de interés. **Suponemos de forma inicial distribuciones normales**.

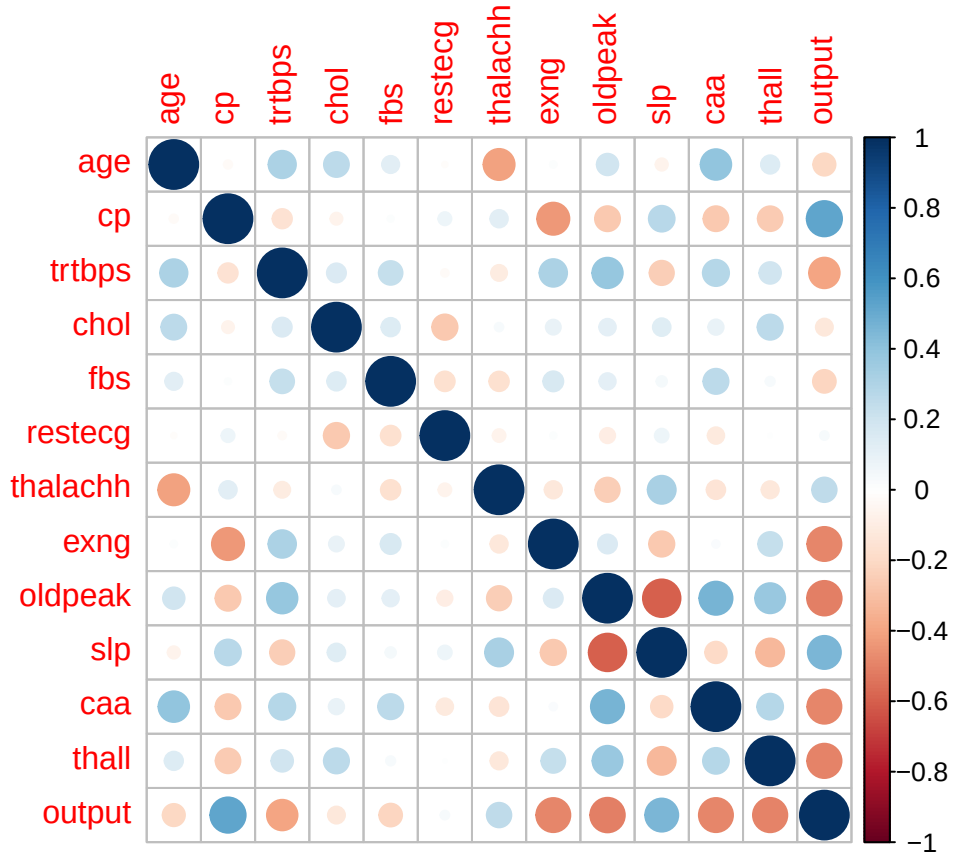
```
datos_cor <- cor(datos)
corrplot(datos_cor, method = "circle")
```



En este caso con el *dataset* completo comprobamos que las variables que tienen una relación más fuerte con *output* son positivamente **cp, thalach y slp** y negativamente **exng, oldpeak, caa, thall y sex**. Las variables restantes incluidas **age** parecen no tener una relación importante con **output**.

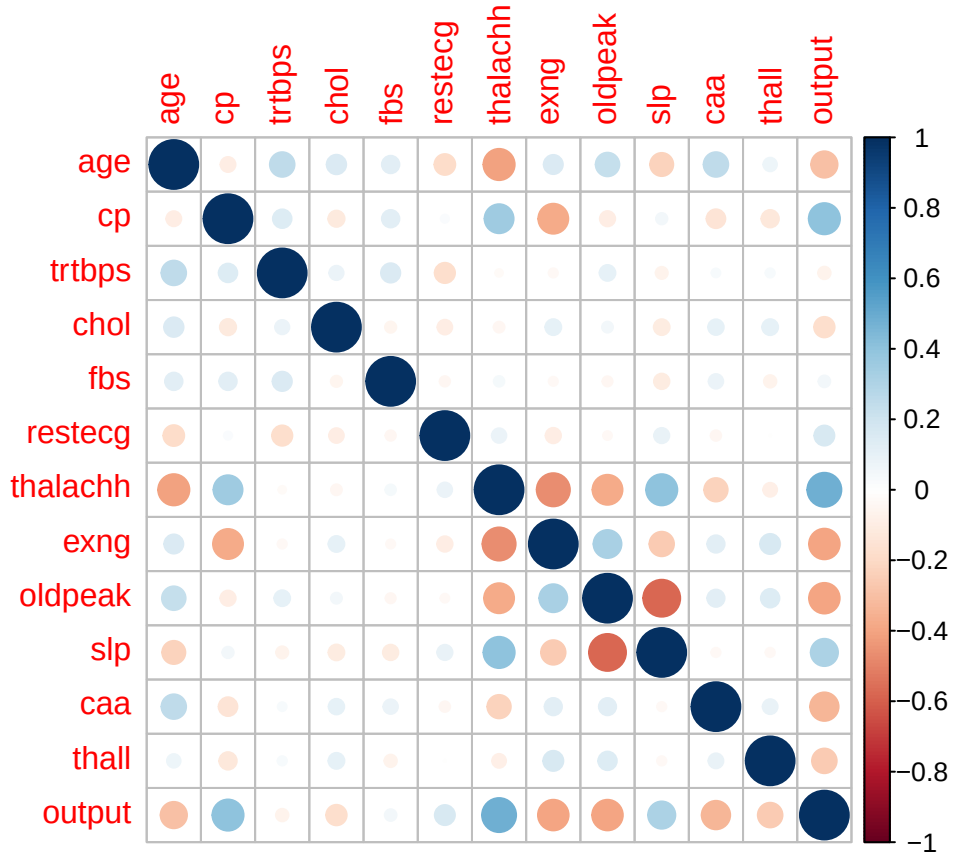
Vamos a realizar las correlaciones en base a la variable *sex* y ver si los resultados obtenidos son similares.

```r
# Separamos mujeres
datos_f <- datos[which(datos$sex==0),]
datos_f[, "sex"] <- list(NULL)
datos_cor_f <- cor(datos_f)
corrplot(datos_cor_f, method = "circle")
```



Con los datos sólo de mujeres las variables con mayor relación positiva son **cp y slp** y **thalach** bastante menos, y relación negativa **exng, oldpeak, caa y thall**, pero también aparece **trtbps** con una fuerte relación negativa.

```r
# Separamos hombres
datos_m <- datos[which(datos$sex==1),]
datos_m[, "sex"] <- list(NULL)
datos_cor_m <- cor(datos_m)
corrplot(datos_cor_m, method = "circle")
```



Con los datos sólo de los hombres las variables con mayor relación positiva son **cp y thalach** y **slp** bastante menos, y relación negativa **exng, oldpeak, caa y thall**, pero también aparece **age** con una fuerte relación negativa.

Como conclusiones:

- Entre el grupo de mujeres y hombres hay alguna diferencia en cuanto a las variables con las que tienen mayor relación principalmente, en el grupo de las mujeres la variable **trtbps** presión arterial en reposo es importante y en el grupo de los hombres **age** influye.
- Hay variables como **restecg, fbs y chol** que parece no tener relación con la variable **output** y podemos eliminarlas del *dataset*.

```
# Eliminamos variables
datos[, c("chol","fbs","restecg")] <- list(NULL)
head(datos)
```

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

```
datos_m[, c("chol","fbs","restecg")] <- list(NULL)
head(datos_m)
```

```
##   age cp trtbps thalachh exng oldpeak slp caa thall output
## 1  63  3    145      150    0     2.3   0   0     1      1
## 2  37  2    130      187    0     3.5   0   0     2      1
## 4  56  1    120      178    0     0.8   2   0     2      1
## 6  57  0    140      148    0     0.4   1   0     1      1
## 8  44  1    120      173    0     0.0   2   0     3      1
## 9  52  2    172      162    0     0.5   2   0     3      1
```

```
datos_f[, c("chol","fbs","restecg")] <- list(NULL)
head(datos_f)
```

```
##    age cp trtbps thalachh exng oldpeak slp caa thall output
## 3   41  1    130      172    0     1.4   2   0     2      1
## 5   57  0    120      163    1     0.6   2   0     2      1
## 7   56  1    140      153    0     1.3   1   0     2      1
## 12  48  2    130      139    0     0.2   2   0     2      1
## 15  58  3    150      162    0     1.0   2   0     2      1
## 16  50  2    120      158    0     1.6   1   0     2      1
```

Ahora vamos a convertir las variables a los tipos más adecuados para el posterior tratamiento de los datos. Todas las variables cualitativas o categóricas las convertimos en variables tipo *factor*. Almacenamos los datos resultantes en un archivo llamado **heart_out.csv**

```
# dataset completo
datos$sex <- as.factor(datos$sex)
datos$output <- as.factor(datos$output)
datos$cp <- as.factor(datos$cp)
datos$exng <- as.factor(datos$exng)
# Sólo mujeres
datos_m$output <- as.factor(datos_m$output)
datos_m$cp <- as.factor(datos_m$cp)
datos_m$exng <- as.factor(datos_m$exng)
# Sólo hombres
datos_f$output <- as.factor(datos_f$output)
datos_f$cp <- as.factor(datos_f$cp)
datos_f$exng <- as.factor(datos_f$exng)
sapply(datos, summary) %>% pander()
```

- **age**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
29	47.5	55	54.37	61	77

- **sex**:

0	1
96	207

- **cp**:

0	1	2	3
143	50	87	23

- **trtbps**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
94	120	130	131.6	140	200

- **thalachh**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
71	133.5	153	149.6	166	202

- **exng**:

0	1
204	99

- **oldpeak**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0.8	1.04	1.6	6.2

- **slp**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	1	1	1.399	2	2

- **caa**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	0	0.7294	1	4

- **thall**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	2	2	2.314	3	3

- **output**:

0	1
138	165

```
write.csv(datos, "heart_out.csv", row.names = FALSE)
```

3 Limpieza de datos

En este apartado vamos a analizar si existen valores nulos y vacíos, así como valores extremos.

3.1 Elementos nulos y vacios

```
any(is.na(datos))
```

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

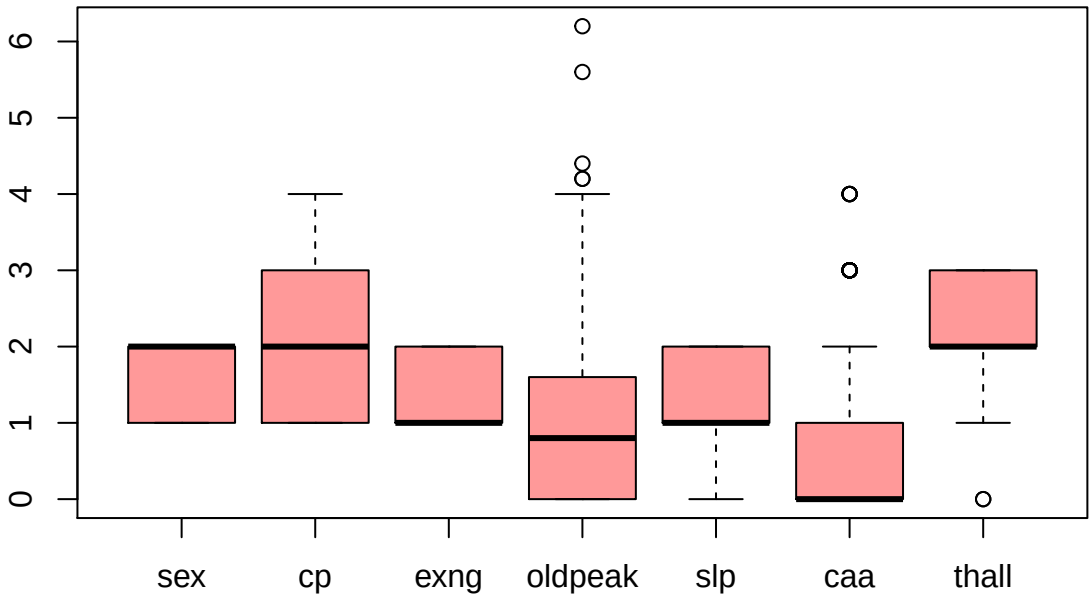
```
apply(is.na(datos),2,mean) %>% pander()
```

age	sex	cp	trtbps	thalachh	exng	oldpeak	slp	caa	thall	output
0	0	0	0	0	0	0	0	0	0	0

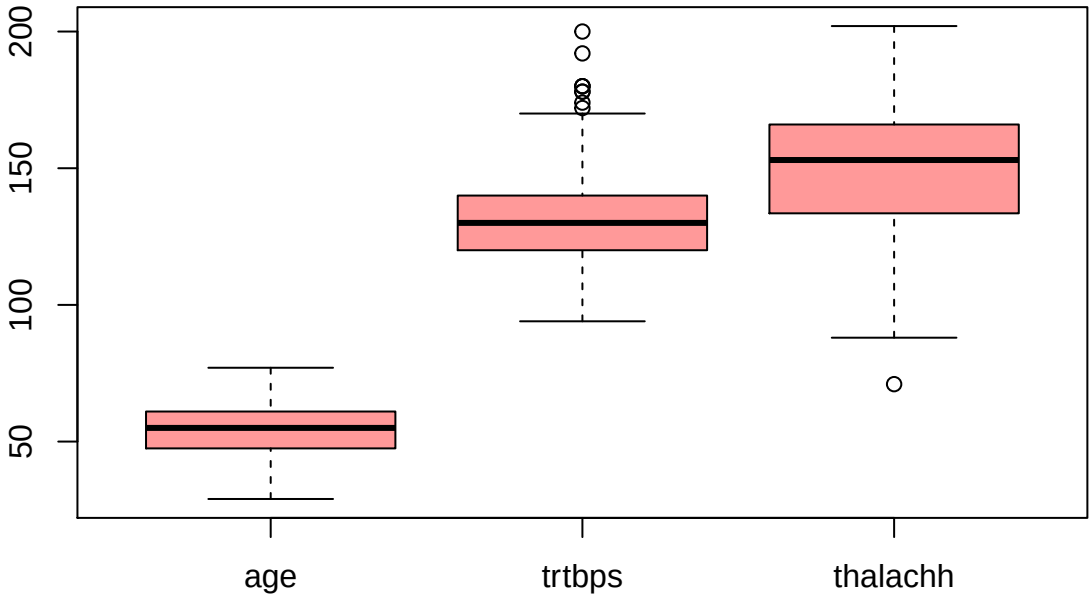
No existen elementos nulos o vacíos en el *dataset*.

3.2 Valores extremos

```
boxplot(datos[,c(2,3,6,7,8,9,10)], col = rgb(1, 0, 0, alpha = 0.4))
```

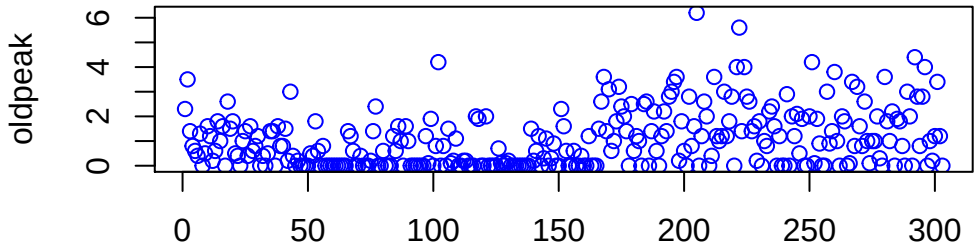



```
boxplot(datos[,c(1,4,5)], col = rgb(1, 0, 0, alpha = 0.4))
```

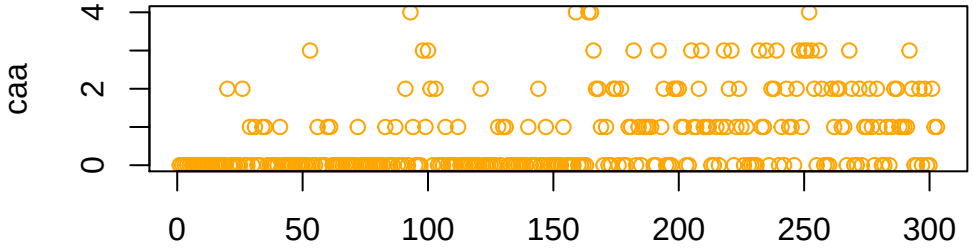


En base a los gráficos anteriores, del primer grupo de variables, **oldpeak, caa y thall** presentan posibles valores extremos o atípicos. Vamos a sacar unos gráficos de dispersión a ver que información nos dan.

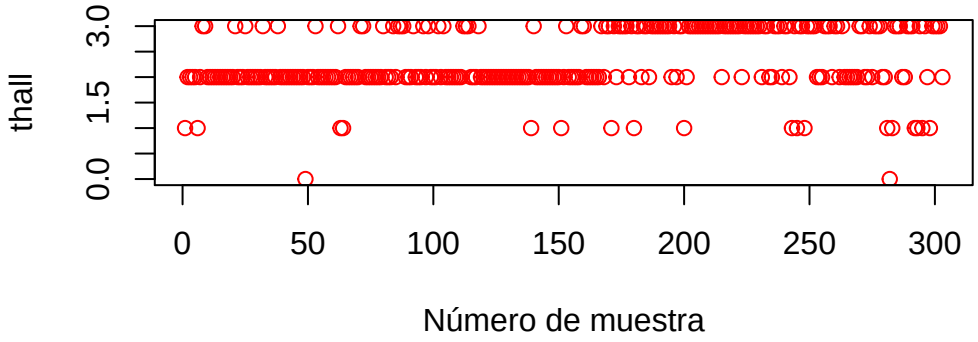
```
plot(datos$oldpeak, col=("blue"), xlab = "Número de muestra", ylab = "oldpeak")
```




```
plot(datos$caa, col=("orange"), xlab = "Número de muestra", ylab = "caa")
```

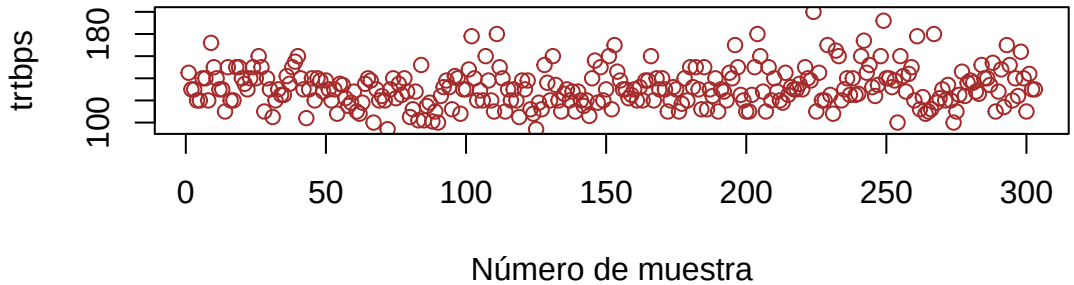



```
plot(datos$thall, col=("red"), xlab = "Número de muestra", ylab = "thall")
```



No parece que haya en este primer grupo de variables datos extremos o atípicos. En el segundo grupo de variables sólo **trtpbs** presenta posibles valores extremos o atípicos, veamos que nos dice el gráfico de dispersión.

```
plot(datos$trtbps, col=("brown"), xlab = "Número de muestra", ylab = "trtbps")
```



En este caso tampoco se puede asegurar que existan valores claramente extremos o atípicos. En base al análisis realizados no hemos detectado valores extremos o atípicos.

4 Análisis de datos

4.1 Selección de grupo de datos

```
# Hasta ahora contamos con los grupos de datos por sexo

# Resumen estadístico del dataset de hombres
summary(datos_m)
```

```
##       age         cp          trtbps         thalachh   exng       oldpeak
##  Min.   :29.00   0:104   Min.   : 94.0   Min.   : 71   0:130   Min.   :0.000
##  1st Qu.:47.00   1: 32   1st Qu.:120.0   1st Qu.:132   1: 77   1st Qu.:0.000
##  Median :54.00   2: 52   Median :130.0   Median :151           Median :0.800
##  Mean   :53.76   3: 19   Mean   :130.9   Mean   :149           Mean   :1.115
##  3rd Qu.:59.50           3rd Qu.:140.0   3rd Qu.:168           3rd Qu.:1.800
##  Max.   :77.00           Max.   :192.0   Max.   :202           Max.   :5.600
##       slp             caa             thall       output
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000   0:114
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1: 93
##  Median :1.000   Median :0.0000   Median :2.000
##  Mean   :1.386   Mean   :0.8116   Mean   :2.401
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :2.000   Max.   :4.0000   Max.   :3.000
```

```
# Resumen estadístico del dataset de mujeres
summary(datos_f)
```

```
##       age         cp          trtbps         thalachh     exng       oldpeak
##  Min.   :34.00   0:39   Min.   : 94.0   Min.   : 96.0   0:74   Min.   :0.000
##  1st Qu.:49.75   1:18   1st Qu.:120.0   1st Qu.:141.2   1:22   1st Qu.:0.000
##  Median :57.00   2:35   Median :131.0   Median :157.0          Median :0.600
##  Mean   :55.68   3: 4   Mean   :133.1   Mean   :151.1          Mean   :0.876
##  3rd Qu.:63.00          3rd Qu.:140.0   3rd Qu.:165.0          3rd Qu.:1.400
##  Max.   :76.00          Max.   :200.0   Max.   :192.0          Max.   :6.200
##       slp             caa             thall       output
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000   0:24
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1:72
##  Median :1.000   Median :0.0000   Median :2.000
##  Mean   :1.427   Mean   :0.5521   Mean   :2.125
##  3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :3.0000   Max.   :3.000
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza

```
# Comprobación de la normalidad en el grupo de datos masculinos (age)
shapiro.test(datos_m$age)
```

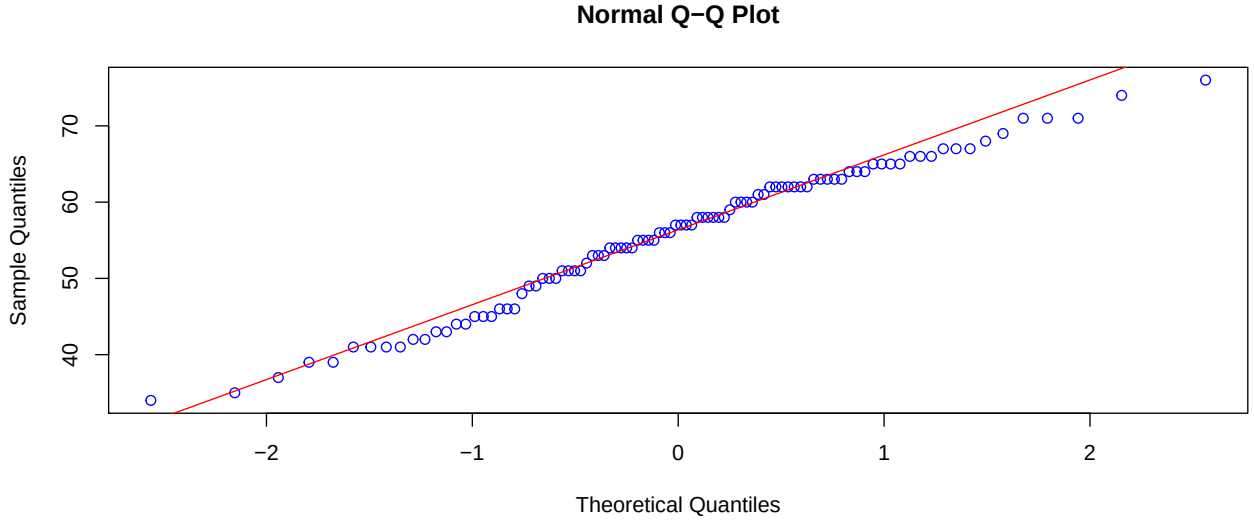
```
##
##  Shapiro-Wilk normality test
##
## data:  datos_m$age
## W = 0.9861, p-value = 0.04
```

```
# Comprobación de la normalidad en el grupo de datos femeninos (age)
shapiro.test(datos_f$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos_f$age
## W = 0.97953, p-value = 0.1386
```

```
# Gráfico Q-Q plot para el grupo de datos masculinos (age)
qqnorm(datos_m$age, col= "orange")
qqline(datos_m$age, col="red")
```


```
# Gráfico Q-Q plot para el grupo de datos femeninos (age)
qqnorm(datos_f$age, col="blue")
qqline(datos_f$age, col="red")
```




```
# Cargar el paquete 'car'
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

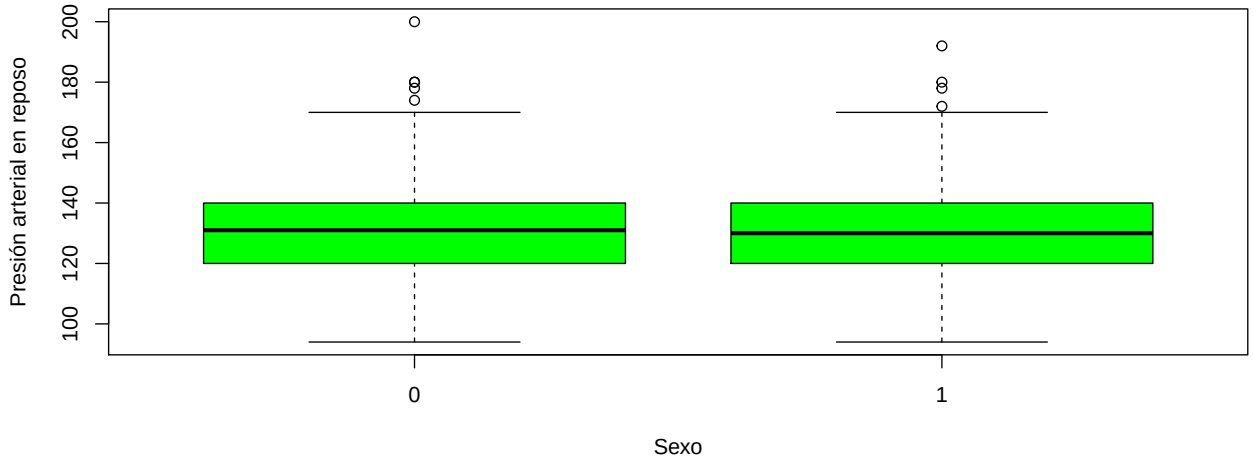
```
## The following object is masked from 'package:dplyr':
```

```
##
##     recode
```

```
# Realizar la prueba de homogeneidad de varianza utilizando leveneTest
leveneTest(trtbps ~ sex, data = datos)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.3593 0.2446
##       301
```

```
# Gráfico de dispersión para comparar la presión arterial en reposo entre los
#grupos masculino y femenino
plot(trtbps ~ sex, data = datos, xlab = "Sexo", ylab = "Presión arterial en reposo"
     , col="green")
```



Para el grupo de datos masculinos (datos_m$age), el resultado de la prueba de normalidad Shapiro-Wilk muestra un valor de W = 0.9861 y un valor de p = 0.04. Dado que el valor de p (0.04) es menor que el nivel de significancia comúnmente utilizado de 0.05, podemos rechazar la hipótesis nula de normalidad. Esto indica que la distribución de las edades en el grupo de datos masculinos no sigue una distribución normal.

Por otro lado, para el grupo de datos femeninos (datos_f$age), el resultado de la prueba de normalidad Shapiro-Wilk muestra un valor de W = 0.97953 y un valor de p = 0.1386. En este caso, el valor de p (0.1386) es mayor que el nivel de significancia de 0.05, lo que indica que no tenemos suficiente evidencia para rechazar la hipótesis nula de normalidad. Esto sugiere que la distribución de las edades en el grupo de datos femeninos se ajusta a una distribución normal.

Por otro lado, el valor de la estadística de prueba F es 1.3593 y el valor p correspondiente es 0.2446. El valor p es la probabilidad de obtener un resultado igual o más extremo que el observado, bajo la suposición nula de que las varianzas de los grupos son iguales.

En este caso, como el valor p (0.2446) es mayor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05), no hay evidencia suficiente para rechazar la hipótesis nula. Por lo tanto, se puede concluir que no hay diferencias significativas en la homogeneidad de la varianza de la presión arterial en reposo entre los grupos de sexo masculino y femenino en el conjunto de datos analizado.

4.3 Pruebas estadísticas

```
# Comparamos la edad promedio entre hombres y mujeres
t.test(datos_m$age, datos_f$age)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  datos_m$age and datos_f$age
## t = -1.6805, df = 175.92, p-value = 0.09464
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.1718589  0.3346005
## sample estimates:
## mean of x mean of y
##  53.75845  55.67708
```

```
# ANOVA para comparar la presión arterial en reposo entre los diferentes tipos de dolor en el pecho
modelo_anova <- aov(trtbps ~ cp, data = datos)
summary(modelo_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## cp            3   2643   881.0   2.919 0.0344 *
## Residuals   299  90248   301.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Correlación entre la edad y la tensión arterial en hombres
cor.test(datos_m$age, datos_m$trtbps)
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos_m$age and datos_m$trtbps
## t = 3.7631, df = 205, p-value = 0.000219
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1220575 0.3774812
## sample estimates:
##       cor
## 0.2541964
```

```
# Prueba de chi-cuadrado para examinar la relación entre el valor máximo de pulsaciones cardíacas
#detectadas y el valor de ST depresión, inducida por el ejercicio relativo al reposo
tabla_chi <- table(datos_f$thalach, datos_f$oldpeak)
chisq.test(tabla_chi)
```

```
## Warning in chisq.test(tabla_chi): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tabla_chi
## X-squared = 1496.4, df = 1248, p-value = 1.372e-06
```

```
# Comparamos la presión arterial en reposo entre hombres y mujeres
wilcox.test(datos_m$trtbps, datos_f$trtbps)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  datos_m$trtbps and datos_f$trtbps
## W = 9285, p-value = 0.3579
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Comparamos el valor máximo de pulsaciones cardíacas entre los diferentes tipos de dolor en el pecho
kruskal.test(thalachh ~ cp, data = datos)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  thalachh by cp
## Kruskal-Wallis chi-squared = 48.216, df = 3, p-value = 1.916e-10
```

```r
# Realizamos un análisis de regresión logística para predecir si un paciente ha sufrido un ataque al
#corazón basado en varias variables
modelo_logistico <- glm(output ~ age + cp + trtbps + thalachh + exng + oldpeak + slp + caa +
                          thall, data = datos, family = "binomial")
summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = output ~ age + cp + trtbps + thalachh + exng +
##     oldpeak + slp + caa + thall, family = "binomial", data = datos)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7040  -0.4441   0.2396   0.5936   2.3838
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.736308   2.270534   0.765 0.444442
## age          0.002770   0.020665   0.134 0.893387
## cp1          1.169514   0.515279   2.270 0.023227 *
## cp2          1.916201   0.430560   4.450 8.57e-06 ***
## cp3          1.814188   0.611110   2.969 0.002991 **
## trtbps      -0.015946   0.009946  -1.603 0.108864
## thalachh     0.017640   0.009530   1.851 0.064185 .
## exng1       -0.889425   0.392757  -2.265 0.023539 *
## oldpeak     -0.567797   0.212190  -2.676 0.007453 **
## slp          0.503225   0.336388   1.496 0.134663
## caa         -0.828189   0.183163  -4.522 6.14e-06 ***
## thall       -1.059247   0.280186  -3.781 0.000157 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 228.61  on 291  degrees of freedom
## AIC: 252.61
##
## Number of Fisher Scoring iterations: 5
```

Dado que el valor p (0.09464) es mayor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05), no hay evidencia suficiente para rechazar la hipótesis nula. Por lo tanto, no se puede concluir que haya una diferencia significativa en la edad promedio entre hombres y mujeres en el conjunto de datos analizado.

El análisis ANOVA muestra que el factor cp tiene un efecto significativo en la presión arterial en reposo, ya que el valor p correspondiente es 0.0344, que es menor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05). Esto significa que hay evidencia suficiente para rechazar la hipótesis nula y concluir que

hay una diferencia significativa en la presión arterial en reposo entre los diferentes tipos de dolor en el pecho.

El coeficiente de correlación de Pearson entre la edad y la tensión arterial en hombres es 0.2541964. El valor p obtenido es 0.000219, que es menor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05). Esto significa que hay evidencia suficiente para rechazar la hipótesis nula de no correlación y concluir que existe una correlación significativa entre la edad y la tensión arterial en hombres.

El intervalo de confianza del 95% para la correlación está entre 0.1220575 y 0.3774812, lo que indica que la correlación verdadera entre la edad y la tensión arterial en hombres está probablemente dentro de este rango.

Se observa un valor de chi-cuadrado de 1496.4 y un p-value de 1.372e-06, que es menor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05). El mensaje de advertencia indica que la aproximación chi-cuadrado puede ser incorrecta debido a las características de los datos.

Esto implica que hay evidencia suficiente para rechazar la hipótesis nula de independencia entre el valor máximo de pulsaciones cardíacas detectadas y el valor de ST depresión, inducida por el ejercicio relativo al reposo en mujeres. En otras palabras, existe una asociación significativa entre estas dos variables en el conjunto de datos analizados.

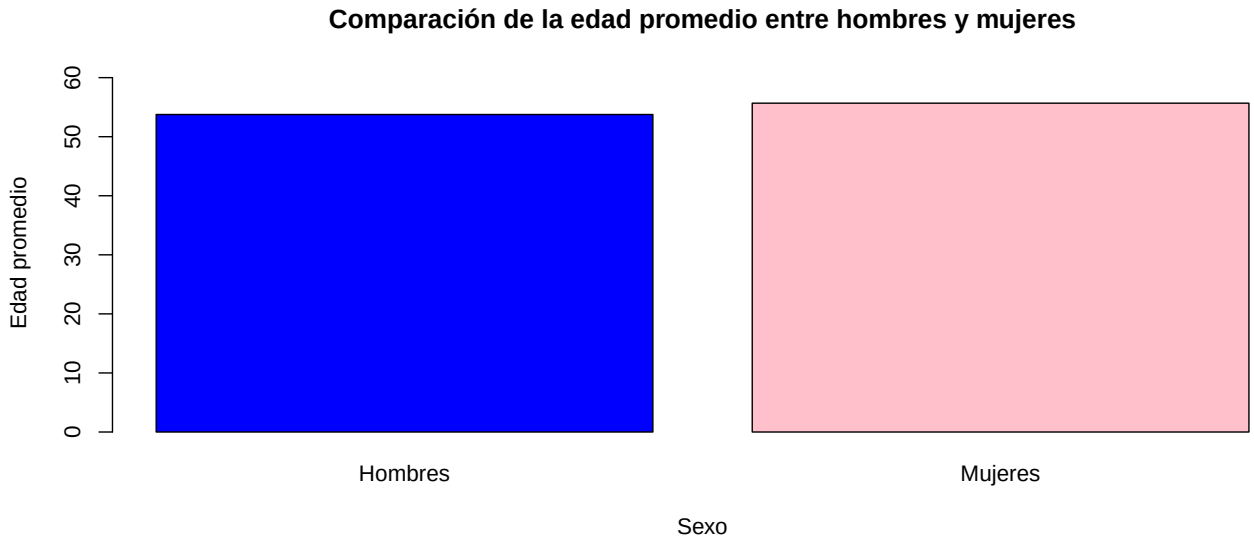
Se observa un valor de W de 9285 y un p-value de 0.3579. Dado que el p-value es mayor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05), no hay suficiente evidencia para rechazar la hipótesis nula de que no hay diferencia en las medianas de la presión arterial en reposo entre hombres y mujeres. Esto implica que no se encontró una diferencia significativa en la presión arterial en reposo entre hombres y mujeres en el conjunto de datos analizados.

Se observa un valor de Kruskal-Wallis chi-squared de 48.216 y un p-value extremadamente pequeño (1.916e-10). Dado que el p-value es mucho menor que el nivel de significancia comúnmente utilizado (por ejemplo, 0.05), hay suficiente evidencia para rechazar la hipótesis nula de que las medianas son iguales entre los diferentes tipos de dolor en el pecho. Esto implica que se encontró una diferencia significativa en el valor máximo de pulsaciones cardíacas entre los diferentes tipos de dolor en el pecho en el conjunto de datos analizados.

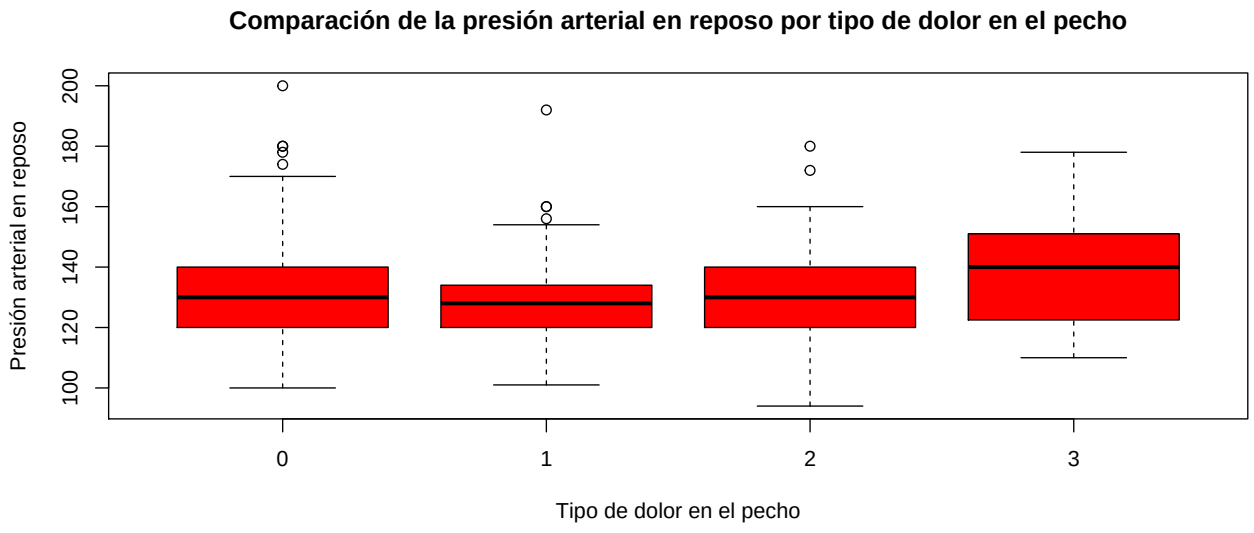
5 Representación de resultados

```r
# Calculamos las medias por grupo
mean_m <- mean(datos_m$age)
mean_f <- mean(datos_f$age)

# Creamos el gráfico de barras
barplot(c(mean_m, mean_f), names.arg = c("Hombres", "Mujeres"),
        ylim = c(0, max(mean_m, mean_f) + 5),
        xlab = "Sexo", ylab = "Edad promedio", col = c("blue", "pink"),
        main = "Comparación de la edad promedio entre hombres y mujeres")
```

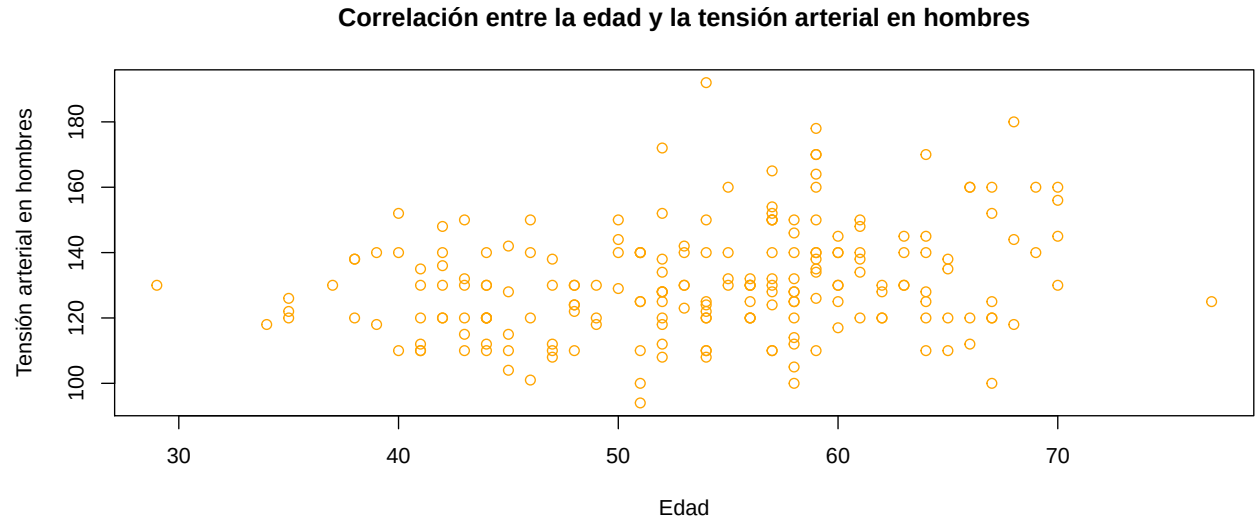



```r
# Creamos el boxplot
boxplot(trtbps ~ cp, data = datos,
        xlab = "Tipo de dolor en el pecho", ylab = "Presión arterial en reposo",
        main = "Comparación de la presión arterial en reposo por tipo de dolor en el pecho",
        col = "red")
```

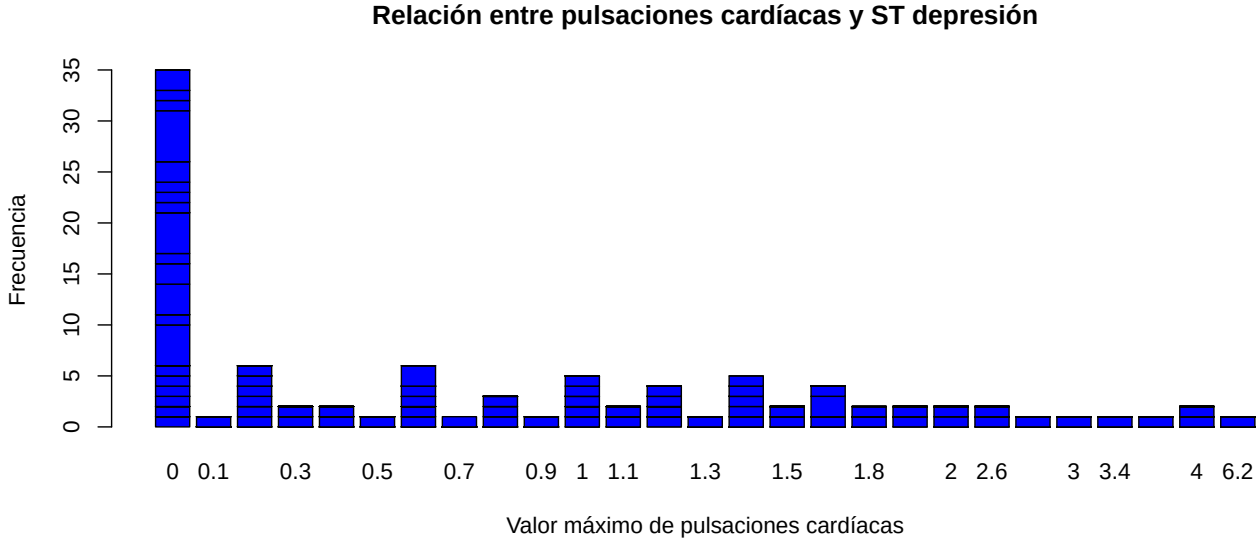



```r
# Creamos el gráfico de dispersión
plot(datos_m$age, datos_m$trtbps,
```

```
     xlab = "Edad", ylab = "Tensión arterial en hombres",
     main = "Correlación entre la edad y la tensión arterial en hombres",
     col = "orange")
```




```
# Creamos el gráfico de barras
barplot(tabla_chi, xlab = "Valor máximo de pulsaciones cardíacas",
        ylab = "Frecuencia", main = "Relación entre pulsaciones cardíacas y ST depresión",
        col = "blue")
```



6 Resolución del problema

Basándonos en el análisis de regresión logística que realizamos previamente, podemos utilizar el modelo logístico para predecir la probabilidad de que una persona sufra un ataque al corazón en función de los valores de las variables independientes.

```
# Obtenemos las predicciones del modelo logístico
predicciones <- predict(modelo_logistico, type = "response", newdata = datos)

# Creamos una columna en el conjunto de datos original con las probabilidades predichas
datos$probabilidad <- predicciones

# Definimos un umbral para determinar si se predice un ataque al corazón o no
umbral <- 0.5

# Agregamos una columna al conjunto de datos que indique si se predice un ataque al corazón o no
datos$prediccion <- ifelse(datos$probabilidad >= umbral, "Sí", "No")

# Visualizar el resultado de la predicción
table(datos$output, datos$prediccion)
```

```
##
##      No  Sí
##   0 107  31
##   1  16 149
```

Con este código, hemos agregado dos columnas nuevas al conjunto de datos: “probabilidad”, que representa la probabilidad predicha de sufrir un ataque al corazón para cada individuo, y “prediccion”, que indica si se predice un ataque al corazón o no en base al umbral establecido.

La tabla resultante muestra la comparación entre los valores reales de la variable “output” y las predicciones realizadas por el modelo. Puedes evaluar la precisión de las predicciones y determinar la tasa de aciertos y errores del modelo.

7 Código

El código se ha realizado en R y se encuentra disponible en el repositorio GitHub https://github.com/fgarciagonzalez2/TD_Practica2.git

8 Vídeo

Los vídeos han sido subidos al aula virtual por cada uno de los alumnos.

Cuadro 13: Tabla de contribuciones

Contribuciones	Firma
Investigación previa	AMV, FEGG
Redacción de las respuestas	AMV, FEGG
Desarrollo del código	AMV, FEGG
Participación en el vídeo	AMV, FEGG