Tipología del dato: PRA2

Autores: Juan María Guerrero y José Ángel Rodríguez

enero 2023

Contents

Antes de comenzar, importamos las librerías dplry y ggplot2 entre otras que podremos necesitar más adelante:

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if(!require('Rmisc')) install.packages('Rmisc'); library('Rmisc')
if(!require('dplyr')) install.packages('dplyr'); library('dplyr')
if(!require('xfun')) install.packages('xfun'); library('xfun')
if (!require('tidyverse')) install.packages('tidyverse'); library('tidyverse')
if(!require("corrplot")) install.packages("corrplot"); library("corrplot")
if (!require('factoextra')) install.packages('factoextra'); library('factoextra')
if (!require('car')) install.packages('car'); library('car')
```

1 Descripción del dataset

El dataset que hemos seleccionado es el propuesto por el propio enunciado de la asignatura, encontrándose este en el enlace https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=download de kaggle, y dentro de este podremos encontrar información relativa a las condiciones de salud de diversas personas que, en mayor o menor riesgo, podrían ser propensas a padecer un infarto debido a dichas condiciones. El objetivo de nuestro análisis será realizar un estudio de los datos para determinar la posible relación de ciertas variables referentes a la salud cardiovascular con el hecho de sufrir un infarto. Entre los posibles casos de análisis que nos ofrece el dataset, vamos a intentar dar respuesta a dos cuestiones principalmente:

- ¿Hay evidencias de que la edad está relacionada con el riesgo de sufrir un infarto?
- ¿Existe una mayor probabilidad de sufrir un infarto siendo hombre?

El fin último de nuestro análisis será determinar cuáles son las variables más determinantes en este aspecto y si, en base a estas, seríamos capaces o no de predecir si alguien está en riesgo alto o no de sufrir un infarto.

En primer lugar, vamos a cargar el fichero csv como se hacía en el apartado anterior y vamos a analizar su estructura:

```
path = 'heart.csv'
df <- read.csv(path)
```

Y vemos a continuación la estructura del mismo:

Verificamos la estructura del juego de datos principal. Vemos el número de columnas que tenemos y ejemplos de los contenidos de las filas.

```
# Reejecutar manualmente esta celda si da problemas.
str(df)
```

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

```
summary(df)
```

```
##       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
```

Como podemos ver el dataset se compone de 303 registros, organizados en 14 columnas. Dentro de este, se aprecian 13 variables con valores enteros y una con valores decimales. Entre las variables con valores enteros podemos ver que 4 de ellas (sex, exng, fbs y output) serán binarias, mientras 5 de ellas (cp, restecg, slp, caa y thall) harán referencia a una característica *cualitativa* en función del valor que tomen.

A continuación vamos a ver a qué hacen referencia las variables propuestas en el dataset. Como apoyo para esto se ha utilizado la fuente https://www.ijrte.org/wp-content/uploads/papers/v8i2S3/B11630782S319.pdf , ya que la información proporcionada en kaggle no estaba actualizada. Como contexto, cabe destacar también que el segmento ST hace referencia a la curva mostrada en el electrocardiograma cuando se contrae el ventrículo, pero no hay electricidad fluyendo a través de él.

VARIABLES A ESTUDIAR

- **age** Edad del paciente
- **sex** Sexo del paciente
- **exng** Angina inducida por ejercicio (1 = sí; 0 = no)
- **caa** Número de vasos principales (0-4)
- **cp** Tipo de dolor torácico: - Valor 0: angina típica - Valor 1: angina atípica - Valor 2: dolor no anginoso - Valor 3: asintomático
- **trtbps** Presión arterial en reposo (en mm Hg)
- **chol** Colesterol en mg/dl obtenido a través del sensor de IMC
- **fbs** Glucemia en ayunas > 120 mg/dl (1 = verdadero; 0 = falso)
- **restecg** Resultados electrocardiográficos en reposo - Valor 0: normal - Valor 1: con anomalía de la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST de > 0,05 mV) - Valor 2: hipertrofia ventricular izquierda probable o definida según los criterios de Estes
- **thalachh** Frecuencia cardiaca máxima alcanzada
- **oldpeak** Valor del último pico en la curva ST
- **slp** Pendiente del pico del segmento ST durante el ejercicio (slope)
- **thall** Trastorno en la sangre conocido como talasemia que hace que no produzcamos suficiente hemoglobina y toma valores: - Valor 0: Heredado de ambos - Valor 1: Heredado del padre - Valor 2: Heredado de la madre - Valor 3: No tiene
- **output** 0 = Probabilidad baja de infarto, 1 = Probabilidad alta de infarto

2 Limpieza de datos

2.1 Existencia de valores nulos, ceros o elementos vacíos:

Antes de continuar vamos a comprobar la existencia de 0s (que presuponemos existirán ya que el 0 se encuentra entre los valores aceptados para determinadas variables):

```
apply(df, 2, function(x) sum(x == 0))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0       96      143        0        0      258      147        0
##     exng  oldpeak      slp      caa    thall   output
##      204       99       21      175        2      138
```

Como podemos ver, aquellas variables que hacen referencia a características tendrán el 0 como un valor válido. Para que nuestro dataframe tenga una aplicación más sencilla a la hora de usar en él un posible modelo predictivo, convertiremos las variables binarias en variables booleanas (TRUE y FALSE) y no eliminaremos

las variables numéricas, si no que simplemente vamos a sumarle uno a las variables que cuenten con un 0 para eliminar la existencia de estos.

```
df$fbs <- as.logical(df$fbs)
df$exng <- as.logical(df$exng)
```

```
df$sex <- as.integer(df$sex + 1)
df$cp <- as.integer(df$cp + 1)
df$restecg <- as.integer(df$restecg + 1)
df$thall <- as.integer(df$thall + 1)
df$oldpeak <- as.integer(df$oldpeak + 1)
df$slp <- as.integer(df$slp + 1)
df$caa <- as.integer(df$caa + 1)
```

Comprobamos ahora que ya no existan ceros y las variables se hayan convertido de forma exitosa:

```
apply(df, 2, function(x) sum(x == 0))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0      258        0        0
##     exng  oldpeak      slp      caa    thall   output
##      204        0        0        0        0      138
```

```
estructura = str(df)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : int  4 3 2 2 1 1 2 2 3 3 ...
##  $ 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     : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ restecg : int  1 2 1 2 2 2 1 2 2 2 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ oldpeak : int  3 4 2 1 1 1 2 1 1 2 ...
##  $ slp     : int  1 1 3 3 3 2 2 3 3 3 ...
##  $ caa     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ thall   : int  2 3 3 3 3 2 3 4 4 3 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Podemos concluir que los cambios se han llevado a cabo correctamente, ya que los valores que identifica como ceros serán los booleanos ya convertidos y la variable objetivo que se construye en base a una probabilidad por lo que tiene sentido mantenerla como 0 y 1.

Acto seguido, veremos las estadísticas básicas y comprobaremos si cuenta con valores nulos y/o vacíos:

```
colSums(is.na(df))
```

```
##      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
```

Como se puede apreciar, no aparecerán valores nulos, vamos a comprobar ahora si existen vacíos:

```
colSums(df==" ")
```

```
##      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
```

```
colSums(df=="")
```

```
##      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
```

De nuevo, podemos ver que no existirán valores vacíos de forma que los datos a primera vista serán correctos.

2.2 Identificación y gestión de valores extremos

Una vez visto esto, vamos ahora a analizar la existencia de valores extremos que se encuentren fuera de los rangos aceptables o que por su valor podamos identificar como erróneos por un posible problema de mala medición al tomar la muestra.

En primer lugar nos fijamos de nuevo en el resumen de las variables de la muestra:

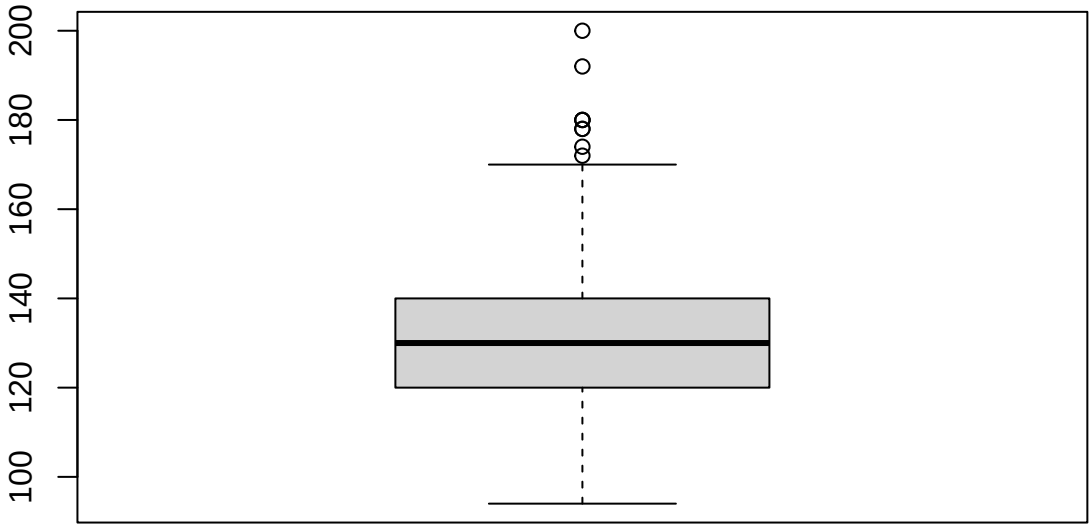
```
summary(df)
```

```
##       age             sex              cp            trtbps     
##  Min.   :29.00   Min.   :1.000   Min.   :1.000   Min.   : 94.0  
##  1st Qu.:47.50   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:120.0  
##  Median :55.00   Median :2.000   Median :2.000   Median :130.0  
##  Mean   :54.37   Mean   :1.683   Mean   :1.967   Mean   :131.6  
##  3rd Qu.:61.00   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:140.0  
##  Max.   :77.00   Max.   :2.000   Max.   :4.000   Max.   :200.0  
##       chol          fbs             restecg         thalachh    
##  Min.   :126.0   Mode :logical   Min.   :1.000   Min.   : 71.0  
##  1st Qu.:211.0   FALSE:258       1st Qu.:1.000   1st Qu.:133.5  
##  Median :240.0   TRUE :45        Median :2.000   Median :153.0  
##  Mean   :246.3                   Mean   :1.528   Mean   :149.6  
##  3rd Qu.:274.5                   3rd Qu.:2.000   3rd Qu.:166.0  
##  Max.   :564.0                   Max.   :3.000   Max.   :202.0  
##     exng            oldpeak           slp             caa        
##  Mode :logical   Min.   :1.000   Min.   :1.000   Min.   :1.000  
##  FALSE:204       1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000  
##  TRUE :99        Median :1.000   Median :2.000   Median :1.000  
##                  Mean   :1.766   Mean   :2.399   Mean   :1.729  
##                  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000  
##                  Max.   :7.000   Max.   :3.000   Max.   :5.000  
##      thall           output     
##  Min.   :1.000   Min.   :0.0000  
##  1st Qu.:3.000   1st Qu.:0.0000  
##  Median :3.000   Median :1.0000  
##  Mean   :3.314   Mean   :0.5446  
##  3rd Qu.:4.000   3rd Qu.:1.0000  
##  Max.   :4.000   Max.   :1.0000
```

Con respecto a la variable decimal oldpeak parece encontrarse en un rango de valores correcto. Como podemos apreciar, todas las variables enteras que toman valores asociados a una característica cualitativa (cp, restecg, slp, caa y thall), y las binarias (sex, exng, fbs y output) tomarán valores dentro de los rangos esperados. De igual forma la variable edad toma valores entre 29 y 77 lo cual parece correcto. Para las variables trtbps, chol y thalachh utilizaremos un boxplot para determinar si existe algún valor que pueda ser un error de medición o que pueda corromper los datos y, de encontrarlo, lo eliminaremos del dataset:

En primer lugar analizamos los valores de la presión:

```
boxplot(df$trtbps)
```




```
out <- boxplot.stats(df$trtbps)$out
out_ind <- which(df$trtbps %in% c(out))
df[out_ind, ]$trtbps
```

```
## [1] 172 178 180 180 200 174 192 178 180
```

Vemos que son valores muy elevados pero que podrían ser perfectamente reales, una tensión de 18 es perfectamente factible en personas con la tensión alta dado que los valores promedio se encuentran entorno a los 12 o 13, por lo que alcanzar incluso 20 podría ser algo real para personas en riesgo de sufrir un infarto. Por ello, en este caso no vamos a descartar estos valores para nuestro dataset.

De aquí en adelante, no representaremos gráficamente los valores extremos para no extender innecesariamente la longitud del documento y simplemente analizaremos sus valores.

Vamos ahora a ver los valores del colesterol:

```
out <- boxplot.stats(df$chol)$out
out_ind <- which(df$chol %in% c(out))
df[out_ind, ]$chol
```

```
## [1] 417 564 394 407 409
```

El colesterol puede considerarse como alto (lo cual indica un posible factor de riesgo para sufrir un infarto) a partir de 200, por ello valores por encima de los 400 o incluso los 560 como vemos que se llegan a alcanzar son valores extremadamente anómalos y que podrían llegar a corromper nuestra muestra pudiendo llegar a ser incluso errores de medición. Por ello vamos a eliminarlos de nuestro dataset:

```
df <- df[-out_ind, ]
```

Finalmente analizamos la frecuencia cardiaca máxima:

```
out <- boxplot.stats(df$thalachh)$out
out_ind <- which(df$thalachh %in% c(out))
df[out_ind, ]$thalachh
```

```
## [1] 71
```

El valor de esta suele calcularse restando nuestra edad a 220, por ello, este valor probablemente se trate de un error de medición por ser demasiado bajo pero en cualquier lugar vamos a visualizar la edad de esta persona para estimar el error en la muestra:

```
t <- df[out_ind, ]$age
t
```

```
## [1] 67
```

Por lo que el valor esperado para la muestra debería ser:

```
220-t
```

```
## [1] 153
```

Como podemos ver, estará muy lejos del valor esperado siendo menos de la mitad, lo cual lo hace un valor extremadamente anómalo y muy posiblemente un error de medición. Ante esto, para garantizar la integridad y validez de los datos de nuestra muestra procedemos a eliminar este registro como en el caso anterior.

```
df <- df[-out_ind, ]
```

Como añadido, vamos a agregar una nueva variable *riesgo* que corresponderá con una representación escrita de la información que nos da la variable *output*, siendo alto si esta vale 1 y bajo si esta vale 0. De igual forma añadiremos dos variables de texto, una que haga referencia al sexo de manera escrita y otra que nos indique a qué grupo de edad pertenece cada una de las personas:

```
df$riesgo <- ifelse(df$output == 0, "bajo", "alto")
df$sexo <- ifelse(df$sex == 1, "hombre", "mujer")
df$edad <- as.character(cut(df$age, breaks = c(0,50,60,90), labels = c("Joven","Media","Mayor")))
```

Visualizamos de nuevo la estructura final de nuestro dataset sobre el que llevaremos a continuación el proceso de análisis.

```
str(df)
```

```
## 'data.frame':    297 obs. of  17 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : int  4 3 2 2 1 1 2 2 3 3 ...
##  $ 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     : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ restecg : int  1 2 1 2 2 2 1 2 2 2 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ oldpeak : int  3 4 2 1 1 1 2 1 1 2 ...
##  $ slp     : int  1 1 3 3 3 2 2 3 3 3 ...
##  $ caa     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ thall   : int  2 3 3 3 3 2 3 4 4 3 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ riesgo  : chr  "alto" "alto" "alto" "alto" ...
##  $ sexo    : chr  "mujer" "mujer" "hombre" "mujer" ...
##  $ edad    : chr  "Mayor" "Joven" "Joven" "Media" ...
```

Finalmente, y una vez acaba la limpieza de datos, podemos guardar nuestro nuevo dataset en un nuevo fichero *heart_clean.csv*:

```
write.csv(df, "../heart_clean.csv")
```

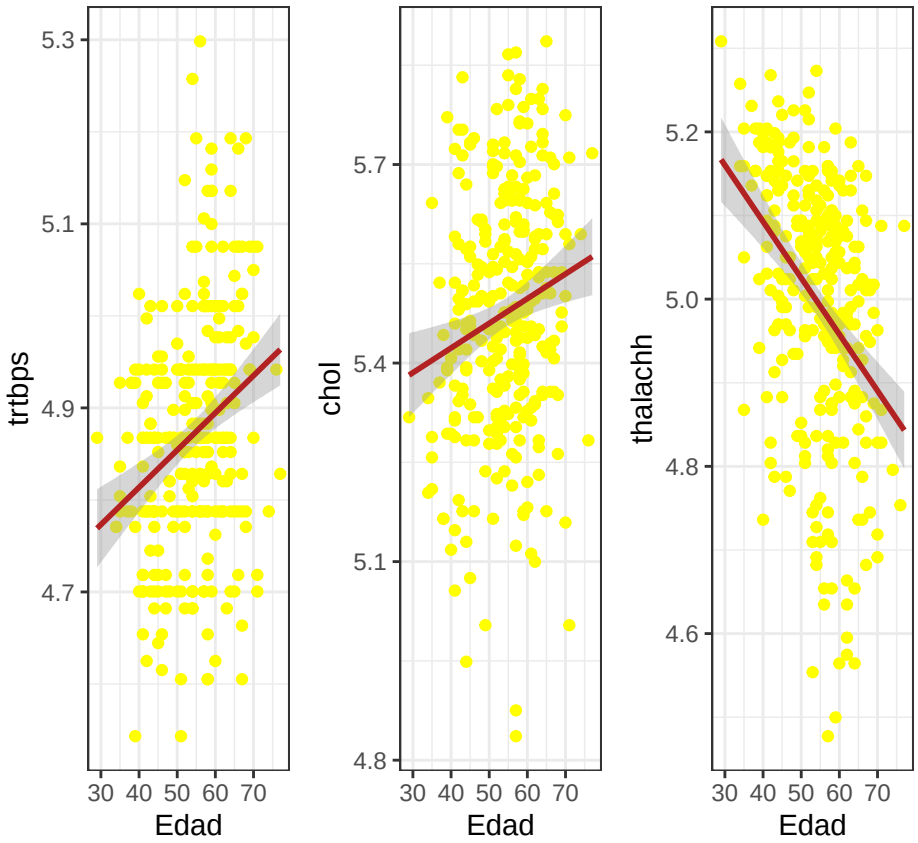
3 Análisis de los datos

3.1 Selección de los grupos de datos a analizar

Vamos antes de nada a comprobar cómo se relacionan las variables numéricas (a priori relacionadas con la posibilidad de sufrir un infarto), con la edad:

```
p = c("trtbps","chol","thalachh")
df_p= df %>% select(all_of(p))
histList2<- vector('list', ncol(df_p))
for(i in seq_along(df_p)){
  message(i)
histList2[[i]]<-local({
  i<-i
  col <-log(df_p[[i]])
  ggp<- ggplot(data = df_p, aes(x = df$age, y=col)) +
    geom_point(color = "yellow") + geom_smooth(method = lm,color = "firebrick") +
    theme_bw() + xlab("Edad") + ylab(names(df_p)[i])
  })

}
multiplot(plotlist = histList2, cols = 4)
```



Obtenemos resultados acorde a lo esperado, viendo que a simple vista tendremos un aumento progresivo de la tensión y el colesterol acorde a nuestra edad, mientras que nuestra frecuencia cardiaca máxima disminuirá claramente con el paso de los años.

Otros dos conjuntos interesantes a la hora de realizar una comparativa podrían ser el de hombres y mujeres, pudiendo realizar una comparativa entre ambos donde podemos ver para cada uno de ellos como se distribuyen los valores de colesterol como representante del riesgo de sufrir un infarto con un rango amplio de valores y, finalmente, ver la relación de estos con la variable objetivo (riesgo real de sufrir un infarto).

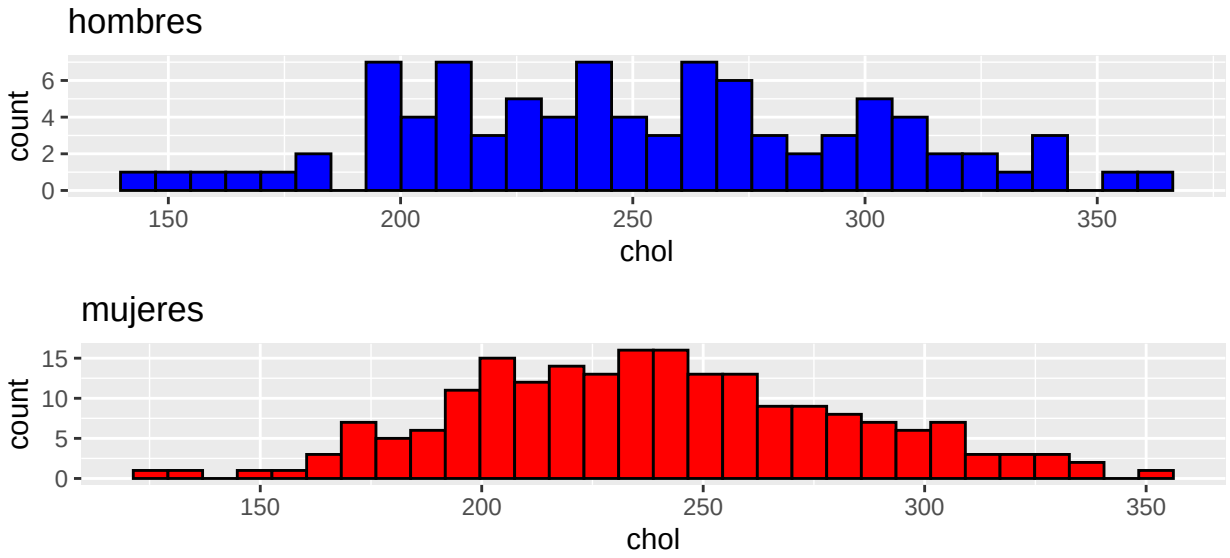
Vamos a analizar en primer lugar la comparativa entre los histogramas de los valores de la tensión seleccionando los grupos de *hombres y mujeres*:

```
df_m <- df %>% filter(df['sex'] == 2)
df_h <- df %>% filter(df['sex'] == 1)
```

Comparamos a continuación los histogramas de los valores del colesterol:

```
r = length(p)


ggp_ch <- ggplot(df_h, aes(x=chol)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  ggtitle("hombres")
ggp_cm <- ggplot(df_m, aes(x=chol)) +
  geom_histogram(bins = 30, fill = "red", color = "black") +
  ggtitle("mujeres")
multiplot(ggp_ch, ggp_cm, coles = 2)
```




```
## [1] 2
```

Como podemos observar, los valores para los hombres parecen ligeramente más desplazados a la derecha en cuestión de colesterol; lo cual podría afectar negativamente.

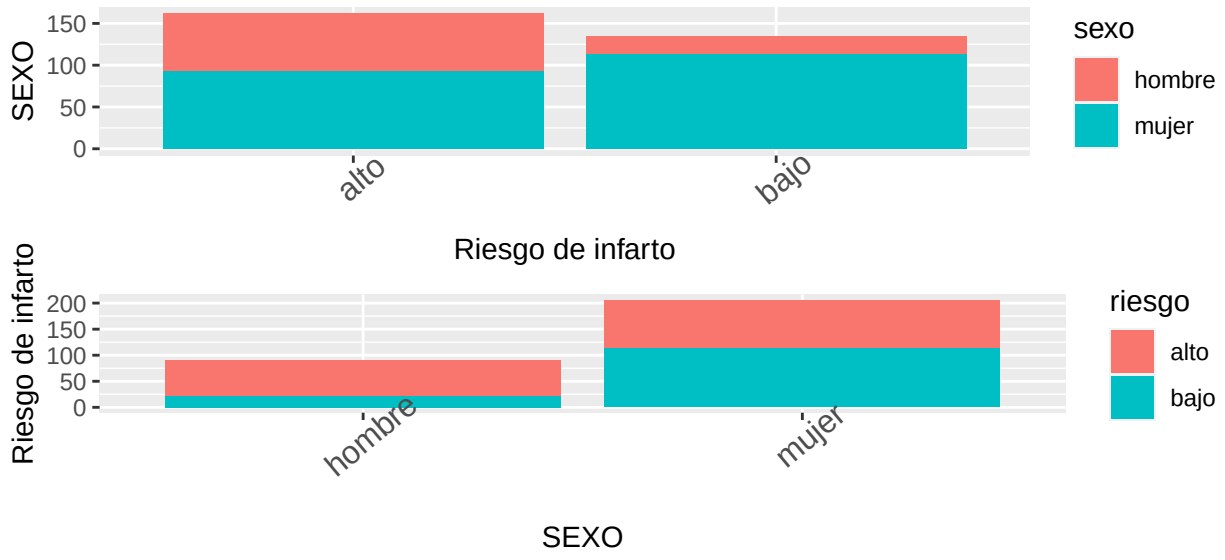
A continuación vamos a visualizar la probabilidad de ser hombre o mujer, en función de si tienes riesgo alto o bajo de sufrir un infarto y, simultáneamente, vamos a visualizar esto mismo, pero analizando para hombres y mujeres cual es el riesgo de tener un infarto según nuestro dataframe:

```
ah1 <- sum(df$output == 1 & df$sex == 1) / sum(df$output == 1) * 100
am1 <- sum(df$output == 1 & df$sex == 2) / sum(df$output == 1) * 100
ah2 <- sum(df$output == 1 & df$sex == 1) / sum(df$sex == 1) * 100
am2 <- sum(df$output == 1 & df$sex == 2) / sum(df$sex == 2) * 100

g1 <- ggplot(data=df,aes(x=riesgo,fill=sexo))+geom_bar()+xlab("Riesgo de infarto")+ylab("SEXO")
g1 <- g1 + theme(axis.text.x = element_text(size=12,angle=40))

g2 <- ggplot(data=df,aes(x=sexo,fill=riesgo))+geom_bar()+xlab("SEXO")+ylab("Riesgo de infarto")
g2 <- g2 + theme(axis.text.x = element_text(size=12,angle=40))

multiplot(g1, g2, coles = 2)
```




```
## [1] 2
```

```
print(sprintf("El %% de gente con riesgo alto de infarto que son mujeres será el %.4f %%",am1))
```

```
## [1] "El % de gente con riesgo alto de infarto que son mujeres será el 57.4074 %"
```

```
print(sprintf("Mientras que el %% de gente con riesgo alto de infarto que son hombres será el %.4f %%",a
```

```
## [1] "Mientras que el % de gente con riesgo alto de infarto que son hombres será el 42.5926 %"
```

```
print(sprintf("El %% de mujeres con riesgo alto de infarto será el %.4f %%",am2))
```

```
## [1] "El % de mujeres con riesgo alto de infarto será el 45.1456 %"
```

```
print(sprintf("Mientras que el %% de hombres con riesgo alto de infarto será el %.4f %%",ah2))
```

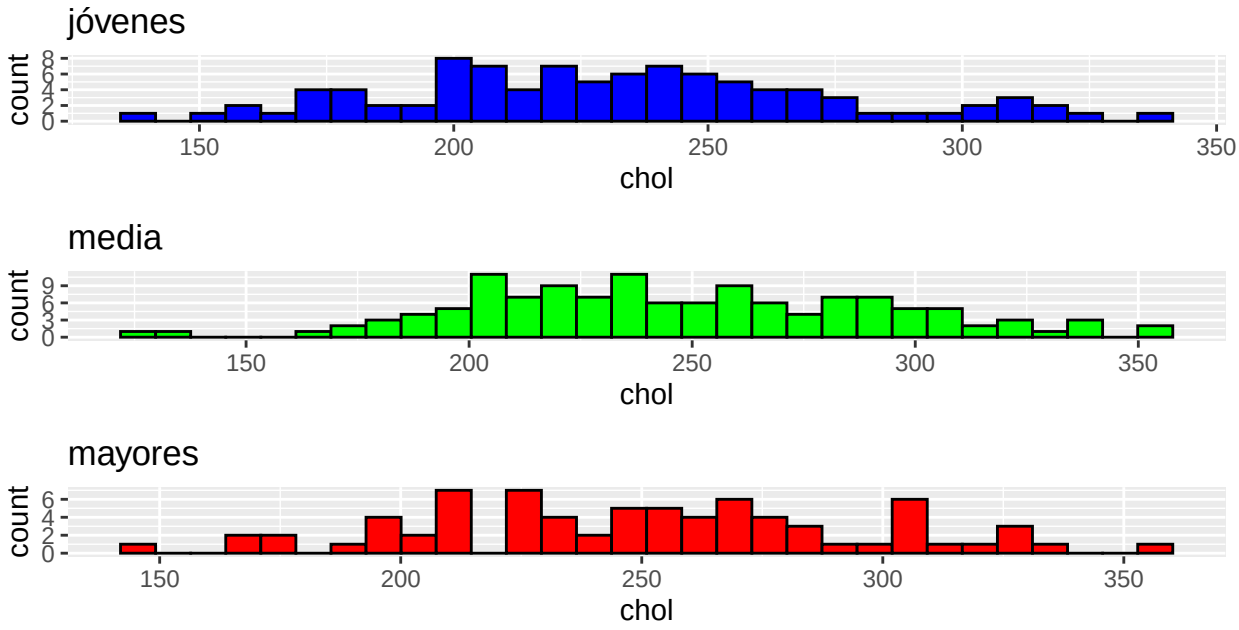
```
## [1] "Mientras que el % de hombres con riesgo alto de infarto será el 75.8242 %"
```

Como podemos ver, dado que por lo observado hay más mujeres en el estudio que hombres, el total de casos de personas con riesgo alto de infarto será mayor en mujeres que en hombres, sin embargo, sin nos fijamos en el riesgo de contraer un infarto diferenciando por sexos, el riesgo para los hombres será del 75% frente al 45% en mujeres, lo cual supone un riesgo mucho mayor.

A continuación analizamos los histogramas de repartición de valores de la variable chol, a priori relacionados con la probabilidad de sufrir un infarto, pero esta vez en lugar de diferenciar por sexo lo haremos diferenciando por *grupos de edad*:

```
df_joven <- df %>% filter(df['edad'] == "Joven")
df_media <- df %>% filter(df['edad'] == "Media")
df_mayor <- df %>% filter(df['edad'] == "Mayor")
```

```
ggp_tj <- ggplot(df_joven, aes(x=chol)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  ggtitle("jóvenes")
ggp_tm <- ggplot(df_media, aes(x=chol)) +
  geom_histogram(bins = 30, fill = "green", color = "black") +
  ggtitle("media")
ggp_tmy <- ggplot(df_mayor, aes(x=chol)) +
  geom_histogram(bins = 30, fill = "red", color = "black") +
  ggtitle("mayores")
multiplot(ggp_tj, ggp_tm, ggp_tmy, coles = 3)
```




```
## [1] 3
```

Aquí podemos encontrar resultados muy variados pero en rasgos generales, podemos confirmar de nuevo que el colesterol seguirá una distribución normal en líneas generales para todos los rangos de edad y, dentro de estos, podemos observar que en los mayores de 60 será donde encontremos valores más dispersos.

3.2 Comprobación de la normalidad y la homogeneidad de la varianza

Un objetivo de la práctica era intentar predecir si alguien está en riesgo alto de sufrir un infarto. Para ello vamos a ver si los grupos analizados anteriormente son susceptibles o no de ser analizados mediante distintos métodos de contraste de hipótesis, regresiones etc

3.2.1 Normalidad

En los histogramas se podía observar una cierta naturaleza normal de los datos de tensión, colesterol y presión máxima. Sin embargo vamos a realizar test estadísticos de comprobación de la normalidad para estas tres

variables.

Para comprobar la normalidad del conjunto vamos a aplicar el test de *Shapito-Wilk*, en este test se asume como hipótesis nula que el conjunto sigue una distribución normal y se calcula el valor del estadístico de prueba y el p-valor. Una vez hecho esto, si el p-valor es mayor que el nivel de significación (en este caso vamos a establecerlo como 0.05) entonces no podremos rechazar la hipótesis nula de que los datos siguen una distribución normal, confirmándose así la normalidad.

En primer lugar, vamos a estudiar la normalidad en la distribución de las mediciones de colesterol en hombres y mujeres, como conjuntos en los que nos basaremos para determinar si es más probable o no que los hombres sufran infartos:

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

```
##
##  Shapiro-Wilk normality test
##
## data:  df$age
## W = 0.98755, p-value = 0.01166
```

```
shapiro.test(df$trtbps)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$trtbps
## W = 0.96522, p-value = 1.447e-06
```

```
shapiro.test(df$chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$chol
## W = 0.993, p-value = 0.1798
```

```
shapiro.test(df$thalachh)
```

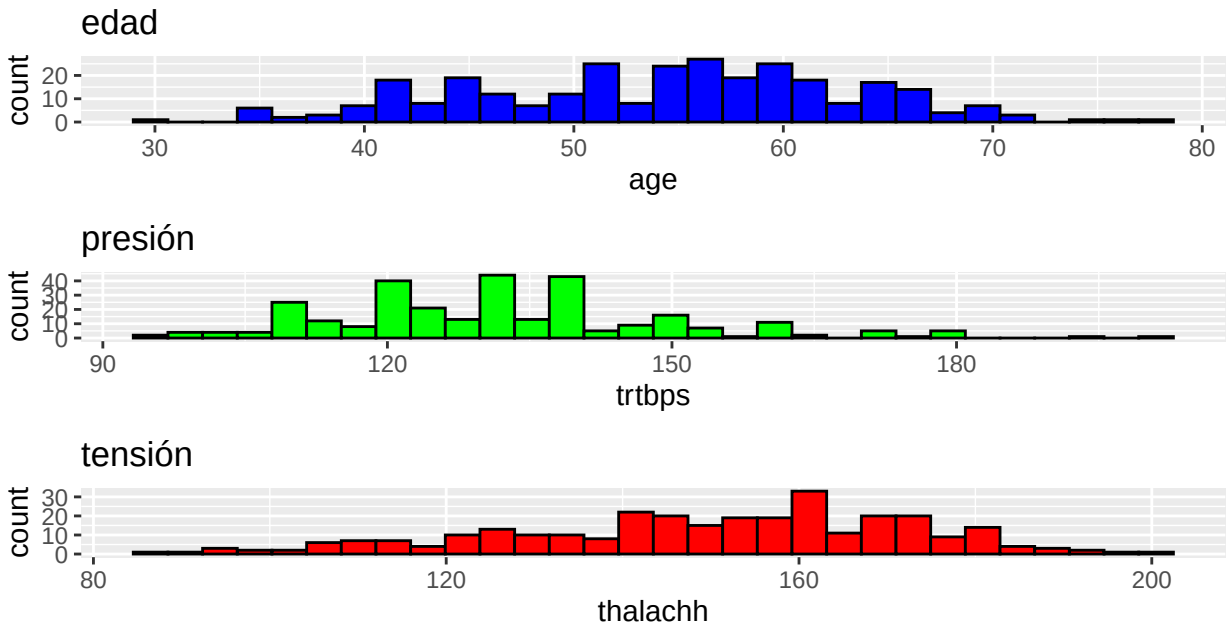
```
##
##  Shapiro-Wilk normality test
##
## data:  df$thalachh
## W = 0.97759, p-value = 0.0001327
```

En los casos de la edad, tensión y la presión máxima, el p-valor está por debajo del nivel de significancia $\alpha = 0.05$“‘. Esto implica que se rechaza la hipótesis nula y no se puede asegurar la normalidad. Sin embargo para el caso del colesterol éste si que tiene un p-valor $> \alpha$ y se acepta la hipótesis nula de ditribución normal de la población.

Vamos, antes de continuar a observar cómo se distribuyen gráficamente estas variables a priori no normales en la muestra:

```
ggp_age <- ggplot(df, aes(x=age)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  ggtitle("edad")
ggp_ten <- ggplot(df, aes(x=trtbps)) +
  geom_histogram(bins = 30, fill = "green", color = "black") +
  ggtitle("presión")
ggp_pre <- ggplot(df, aes(x=thalachh)) +
```

```
  geom_histogram(bins = 30, fill = "red", color = "black") +
  ggtitle("tensión")
multiplot(ggp_age, ggp_ten, ggp_pre, coles = 3)
```




```
## [1] 3
```

Podemos apreciar que visualmente si tendrán una distribución normal, por lo que para corroborar los resultados anteriores probaremos ahora con el test de *Kolmogorov-Smirnov*, buscando obtener una confirmación de lo visto anteriormente:

```
ks.test(df$age, pnorm, mean(df$age), sd(df$age))
```

```
## Warning in ks.test(df$age, pnorm, mean(df$age), sd(df$age)): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df$age
## D = 0.076438, p-value = 0.06219
## alternative hypothesis: two-sided
```

```
ks.test(df$trtbps, pnorm, mean(df$trtbps), sd(df$trtbps))
```

```
## Warning in ks.test(df$trtbps, pnorm, mean(df$trtbps), sd(df$trtbps)): ties
## should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
```

```
##
## data:  df$trtbps
## D = 0.10501, p-value = 0.002859
## alternative hypothesis: two-sided
```

```
ks.test(df$thalachh, pnorm, mean(df$thalachh), sd(df$thalachh))
```

```
## Warning in ks.test(df$thalachh, pnorm, mean(df$thalachh), sd(df$thalachh)): ties
## should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  df$thalachh
## D = 0.072052, p-value = 0.09157
## alternative hypothesis: two-sided
```

En este caso el test de Kolmogorov-Smirnov sí aceptaría la hipótesis de normalidad de la población para la presión máxima y la edad pero no para la tensión cuya normalidad seguimos sin poder asumir.

Igualmente, disponemos de una muestra de casi 300 observaciones y vamos a asumir que por el teorema central del límite las distribución de la media de las dos poblaciones cuya normalidad no aceptaba el test de Saphiro-Wilk sí se acerca cada vez más a una distribución normal.

3.2.2 Homocedasticidad

Comprobamos a continuación la igualdad de varianzas entre las tres variables citadas. Como se planteaba en la descripción del dataset, teníamos dos principales cuestiones que resolver:

- ¿Hay evidencias de que la edad está relacionada con el riesgo de sufrir un infarto?

Para esto, la idea es comparar las poblaciones de la edad y el riesgo de infarto. Comprobamos la homocedasticidad de estas dos variables:

```
leveneTest(age ~ riesgo, data = df)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1  6.8666 0.009237 **
##       295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor 0.009 es inferior al nivel de significancia 0.01. Esto implica que las variables presentan varianzas estadísticamente diferentes y no se pueden aplicar test paramétricos de contrastes de hipotesis.

La otra pregunta a la que queríamos contestar era:

- ¿Existe una mayor probabilidad de sufrir un infarto siendo hombre?

En este caso se aplicará un test estadístico sobre dos variables categóricas, para lo que no hace falta la comprobación de homocedasticidad.

3.3 Pruebas estadísticas para comparación de grupos de datos

- ¿Hay evidencias de que la edad está relacionada con el riesgo de sufrir un infarto?

Como hemos comprobado que no hay igualdad de varianza estadística entre la edad y el riesgo de infarto, vamos a aplicar un test de Wolcoxon y Mann-Whitney, que no es paramétrico.

Las hipótesis serían: - H0: No hay diferencias significativas entre la edad y el riesgo de infarto - H1: Hay diferencias significativas en en riesgo de infarto en base a la edad

```
wilcox.test(age ~ riesgo, data=df)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by riesgo
## W = 7794, p-value = 2.005e-05
## alternative hypothesis: true location shift is not equal to 0
```

En este caso, el p-valor obtenido es menor al nivel de significancia por lo que podemos afirmar que se rechaza la hipótesis nula, confirmando así que *la edad tiene relación con la posibilidad de sufrir un infarto.*

3.3.1 ¿Existe una mayor probabilidad de sufrir un infarto siendo hombre?

En este apartado se realizan los tests necesarios para la comprobación de las siguientes hipótesis:

- H0: No hay relación entre sexos a la hora de sufrir un infarto.
- H1: Hay relación en la probabilidad de sufrir un infarto por ser hombre o mujer.

Para ello se aplica un test de χ^2.

```
chisq.test(df$sexo, df$riesgo)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$sexo and df$riesgo
## X-squared = 22.739, df = 1, p-value = 1.856e-06
```

Se rechaza la hipótesis nula y podemos confirmar que, efectivamente, habrá más riesgo de sufrir un infarto por ser hombre.

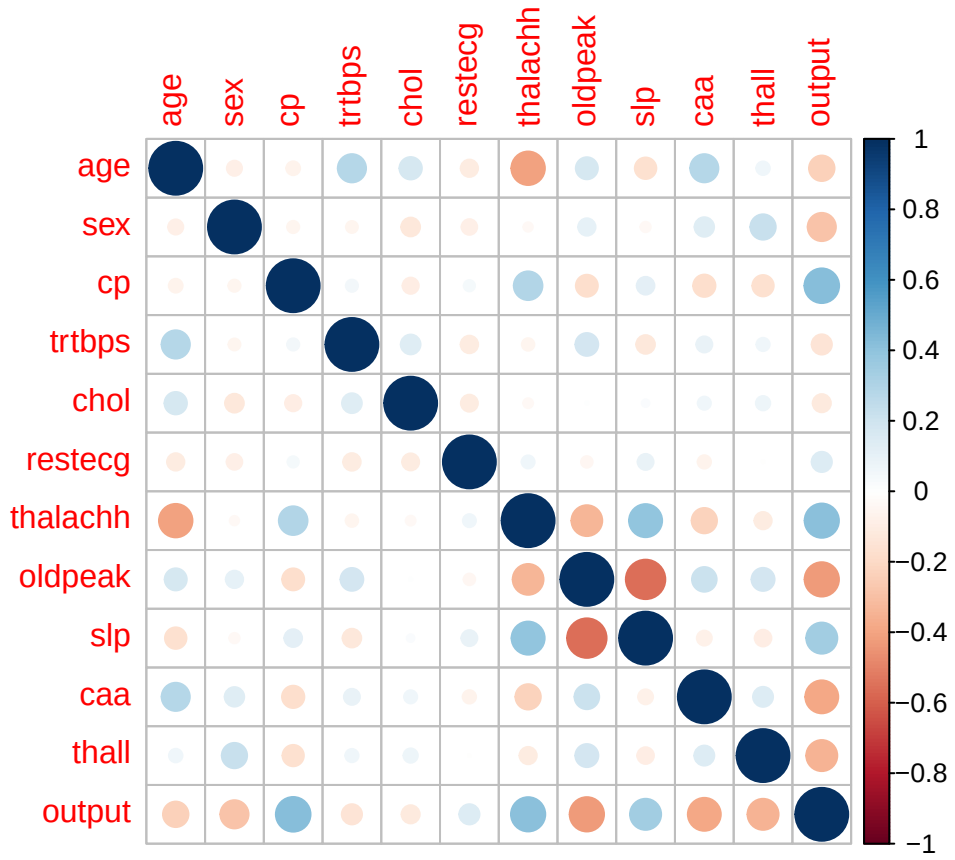
3.3.2 Correlación de variables

Antes de proceder a crear un modelo predictivo con el que poder estimar la probabilidad de sufrir un infarto para nuevos potenciales pacientes, vamos a estudiar la *correlación* de variables para poder determinar cuáles podrían ser las más determinantes a la hora de realizar nuestra predicción.

Antes de calcular la matriz de correlaciones vamos a realizar visualizar de forma gráfica una estimación de la correlación entre variables mediante el método corrplot, donde incluiremos todas las variables:

```
numericas <- c("age", "sex", "cp", "trtbps", "chol", "restecg", "thalachh",
               "oldpeak", "slp", "caa", "thall", "output")
df_num <- df[, numericas]
```

```
options(warn=-1)
corrplot(cor(df_num))
```



Como podemos apreciar, en nuestro dataframe existirán columnas fuertemente relacionadas. Nos interesan sobre todo los resultados de la última fila/ columna de la representación, donde podemos apreciar cómo se relacionan las variables con la probabilidad de sufrir un infarto. A simple vista podemos ver que las variables más relacionadas con este hecho a priori serán la frecuencia cardiaca máxima (thalachh), el tipo de dolor torácico (cp) o el slope (slp) con el hecho de sufrirlo; así como el número de vasos principales (caa), el valor del último pico de la curva ST (oldpeak), o incluso en menor medida el tener un trastorno en la sangre (thall) y la edad (age).

Para corroborar lo visto gráficamente, vamos a obtener la matriz de correlaciones entre estas variables y la probabilidad de sufrir un infarto (output). Para llevarlo a cabo utilizaremos el coeficiente de correlación de *Spearman*, ya que no todas las variables seguirán una distribución normal:

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")


for (i in 1:(ncol(df_num) - 1)) {
  spearman_test = cor.test(df_num[,i],
                           df_num[,length(df_num)],
                           method = "spearman",
                           exact=FALSE)
  corr_coef = spearman_test$estimate
  p_val = spearman_test$p.value

  pair = matrix(ncol = 2, nrow = 1)
  pair[1][1] = corr_coef
  pair[2][1] = p_val
```

```
  corr_matrix <- rbind(corr_matrix, pair)
  rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(df_num)[i]
}
print(corr_matrix)
```

```
##            estimate      p-value
## age      -0.2478974 1.547487e-05
## sex      -0.2840303 6.453497e-07
## cp        0.4557786 1.218889e-16
## trtbps   -0.1223010 3.514156e-02
## chol     -0.1287612 2.649235e-02
## restecg   0.1568579 6.756223e-03
## thalachh  0.4259290 1.613720e-14
## oldpeak  -0.4360428 3.251656e-15
## slp       0.3711750 3.914810e-11
## caa      -0.4570217 9.840155e-17
## thall    -0.4040073 4.346026e-13
```

De nuevo, teniendo en cuenta que los valores más correlacionados con el riesgo de infarto serán aquellos cuyo valor absoluto sea más cercano a 1, podemos confirmar lo visto anteriormente destacando las variables *cp*, *caa*, *oldpeak*, *thalachh*, *thall* por ese orden y, en menor medida, *slp*, *sex* y *age*. También será importante fijarnos en el p-valor que nos dará una estimación del valor estadístico para cada una de las correlaciones obtenidas.

3.3.3 Modelo predictivo

Vamos ahora a intentar construír una regresión logística multivariable como modelo predictivo, basándonos de forma inicial en las variables con un mayor rango de valores que pueden quizá ofrecer más información en base a su valor estadístico por lo visto anteriormente.

```
summary(glm(formula = output ~ trtbps + chol, family = "binomial", data = df))
```

```
##
## Call:
## glm(formula = output ~ trtbps + chol, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5596  -1.2287   0.9079   1.0763   1.5905
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.321728   1.067993   3.110  0.00187 **
## trtbps      -0.015453   0.006930  -2.230  0.02576 *
## chol        -0.004534   0.002662  -1.703  0.08857 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.27  on 296  degrees of freedom
## Residual deviance: 400.03  on 294  degrees of freedom
## AIC: 406.03
##
## Number of Fisher Scoring iterations: 4
```

El criterio de información de Akaike que informa de la bondad del modelo es de 406.03, considerablemente

alto. Vamos a comprobar si añadir variables cuantitativas, a priori relacionadas con la variable objetivo por lo visto anteriormente, pueden mejorar los resultados:

```
summary(glm(formula = output ~ trtbps + chol + restecg + sex + age + thalachh + oldpeak,
            family = "binomial",
            data = df))
```

```
##
## Call:
## glm(formula = output ~ trtbps + chol + restecg + sex + age +
##     thalachh + oldpeak, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3809  -0.7554   0.3310   0.7490   2.8380
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.049420   2.253885   1.353   0.1761
## trtbps      -0.011352   0.009115  -1.245   0.2130
## chol        -0.009034   0.003578  -2.525   0.0116 *
## restecg      0.319814   0.284516   1.124   0.2610
## sex         -1.764271   0.364142  -4.845 1.27e-06 ***
## age         -0.016408   0.018709  -0.877   0.3805
## thalachh     0.037215   0.008082   4.604 4.13e-06 ***
## oldpeak     -0.763589   0.171407  -4.455 8.40e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.27  on 296  degrees of freedom
## Residual deviance: 283.64  on 289  degrees of freedom
## AIC: 299.64
##
## Number of Fisher Scoring iterations: 5
```

Los resultados son mejores, pero los p-valores indican que las variables `oldpeak`, `thalachh` y `sex` no son determinantes en la predicción. Si los eliminamos:

```
summary(glm(formula = output ~ trtbps + chol + restecg + age, family = "binomial", data = df))
```

```
##
## Call:
## glm(formula = output ~ trtbps + chol + restecg + age, family = "binomial",
##     data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.848  -1.122   0.728   1.064   1.677
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.805521   1.284377   2.963  0.00305 **
## trtbps      -0.008489   0.007267  -1.168  0.24277
## chol        -0.003050   0.002744  -1.112  0.26629
```

```
## restecg      0.449225   0.231942   1.937  0.05277 .
## age         -0.044982   0.014548  -3.092  0.00199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.27  on 296  degrees of freedom
## Residual deviance: 385.50  on 292  degrees of freedom
## AIC: 395.5
##
## Number of Fisher Scoring iterations: 4
```

En base al criterio de información de Akaike este sería el mejor modelo predictor en base a una regresión logística.

$$logit(p) = 3.8055 - 0.0085 \times trtbps - 0.0031 \times chol + 0.4492 \times restecg - 0.0450 \times age$$

4 Conclusiones

El set de datos escogido es muy rico en cuanto a información a pesar de no contener un gran volumen de datos, sin embargo este es variado y representativo, por lo que hemos podido dar respuesta a las preguntas planteadas de forma eficiente.

Con respecto al hecho de si `la edad influye o no en el riesgo de padecer un infarto`, podemos determinar que, efectivamente, este `SÍ` será un factor determinante, esto se ha confirmado mediante un test de contraste de hipótesis y, no solo eso, si no que además será un factor determinante a la hora de poder predecir con exactitud si una persona tiene o no un riesgo alto de sufrir un infarto.

Con respecto a si `siendo hombre es más probable sufrir un infarto`, de nuevo se ha confirmado en un test de contrastes unilateral que esto `SÍ` será así. Sin embargo, la variable sexo no será una de las más importantes a la hora de predecir el riesgo de infarto, pues como hemos podido observar en la matriz de correlaciones esta se encontrará en un punto medio donde es importante, pero no será ni de las que mayor relación tengan con la variable objetivo ni de las que mayor significancia estadística tengan, llegando incluso a ser rechazada en nuestro modelo de regresión logística en favor de otras.

Finalmente, se ha podido desarrollar un `modelo de regresión logística eficiente`, con el que poder predecir el riesgo de sufrir un infarto, estimando los valores óptimos del predictor según el criterio de información de Akaine (AIC). Las variables más determinantes para esto han resultado ser la presión (trtbps), el colesterol (chol), los resultados electrocardiográficos en reposo (restecg) y, por último, la edad (age).

5 Contribuciones y firma

Contribuciones	Firma
Investigación previa	Juan María Guerrero Carrasco, José Ángel Rodríguez Murillo
Redacción de las respuestas	Juan María Guerrero Carrasco, José Ángel Rodríguez Murillo
Desarrollo del código	Juan María Guerrero Carrasco, José Ángel Rodríguez Murillo
Participación en el vídeo	Juan María Guerrero Carrasco, José Ángel Rodríguez Murillo