# Tipología y ciclo de vida de los datos: Práctica 2

Autores: David Moliner Mateu y Noel Casado Soler

Enero 2023

## Contents

# 1. Descripción del dataset

Nuestro proyecto consta en el análisis de un dataset que contiene información médica relacionada con el corazón.

Para llevarlo a cabo, necesitamos obtener los datos de una fuente externa a nosotros y por ese motivo recurrimos al repositorio *Kaggle*. Nuestro dataset será solamente uno, el propuesto para esta práctica. Este se encuentra en formato .csv y se llama *Heart Attack Analysis & Prediction Dataset*.

Cita: Rashik Rahman. (Marzo 2021). Heart Attack Analysis & Prediction Dataset. Recuperado [Enero 2023] de https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

Cargamos el dataset y hacemos una copia sobre la cual trabajaremos.

```
heart_original = read.csv(file = "../data/heart.csv", header = TRUE)
heart = heart_original
```

Mostramos los primeros registros para ver qué tipos de valores tenemos y mostramos un resumen estadístico de las variables.

```
head(heart)
```

```
##   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
```

```
summary(heart)
```

```
##       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
```

Únicamente tenemos 303 observaciones, lo cual es un número bastante limitado para obtener conclusiones no sesgadas, y contamos con 14 variables numéricas y categóricas.

Según la información que podemos obtener a través de *Kaggle*, las descripciones de estas 14 variables son las siguientes:

- **age:** edad en años
- **sex:** sexo (hombre = 1, mujer = 0)
- **cp:** tipo de dolor de pecho (asimptomático = 0, angina típica = 1, angina atípica = 2, dolor no relacionado con angina = 3)
- **trtbps:** presión sanguínea en reposo (mm Hg)
- **chol:** colesterol en mg/dl
- **fbs:** azúcar en sangre en ayunas superior a 120 mg/dl (Verdadero = 1, Falso = 0)
- **restecg:** electrocardiograma en reposo (hipertrofia = 0, normal = 1, anormalidad en onda ST-T = 2)
- **thalachh:** máximo ritmo cardíaco obtenido
- **exng:** angina inducida por el ejercicio (Sí = 1, No = 0)
- **oldpeak:** depresión en la onda ST inducida por ejericio
- **slp:** pendiente del pico del segmento ST (pendiente negativa = 0, plana = 1, pendiente positiva = 2)
- **caa:** número de vasos prinicpales
- **thall:** resultado de prueba de esfuerzo nuclear ()
- **output:** predicción de ataque al corazón

## 2. Integración y selección

Hemos decidido realizar una subselección de variables para enfocar nuestro análisis. Dejaremos fuera las variables que están relacionadas con la angina inducida por el ejercicio, por lo que nuestro dataset ahora constará de las siguientes variables:

- **age:** edad en años
- **sex:** sexo (hombre = 1, mujer = 0)
- **cp:** tipo de dolor de pecho (asimptomático = 0, angina típica = 1, angina atípica = 2, dolor no relacionado con angina = 3)
- **trtbps:** presión sanguínea en reposo (mm Hg)
- **chol:** colesterol en mg/dl
- **fbs:** azúcar en sangre en ayunas superior a 120 mg/dl (Verdadero = 1, Falso = 0)
- **restecg:** electrocardiograma en reposo (hipertrofia = 0, normal = 1, anormalidad en onda ST-T = 2)
- **thalachh:** máximo ritmo cardíaco obtenido
- **thall:** resultado de prueba de esfuerzo nuclear ()
- **output:** predicción de ataque al corazón

```
data = heart[,c("age", "sex", "cp", "trtbps", "chol", "fbs", "restecg", "thalachh", "thall", "output")]
```

Convertimos las variables de nuestro subset a numéricas, si son cuantitativas, y a factores, si son categóricas.

```
sapply(data, function(x) class(x))
```

```
##       age       sex        cp    trtbps      chol       fbs   restecg  thalachh
## "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"
##     thall    output
## "integer" "integer"
```

```
data$age <- as.numeric(data$age)
data$sex <- as.factor(data$sex)
data$cp <- as.factor(data$cp)
data$trtbps <- as.numeric(data$trtbps)
data$chol <- as.numeric(data$chol)
data$fbs <- as.factor(data$fbs)
data$restecg <- as.factor(data$restecg)
data$thalachh <- as.numeric(data$thalachh)
data$thall <- as.factor(data$thall)
data$output <- as.factor(data$output)
```

Generamos una nueva variable categórica por tramos de edad para poder usarla como grupo de comparación más adelante.

- Grupo A: 0 a 40 años de edad
- Grupo B: 41 a 60 años de edad
- Grupo C: Mayores de 60 años

```
data$age_group <- cut(data$age, breaks = c(0,40,60,100), labels = c("A","B","C"))
```

Con el nuevo dataset listo, mostramos nuevamente los primeros registros para ver qué tipos de valores tenemos y mostramos el resumen estadístico de las variables.

```
head(data)
```

```
##   age sex cp trtbps chol fbs restecg thalachh thall output age_group
## 1  63   1  3    145  233   1       0      150     1      1         C
## 2  37   1  2    130  250   0       1      187     2      1         A
## 3  41   0  1    130  204   0       0      172     2      1         B
## 4  56   1  1    120  236   0       1      178     2      1         B
## 5  57   0  0    120  354   0       1      163     2      1         B
## 6  57   1  0    140  192   0       1      148     1      1         B
```

```
summary(data)
```

```
##       age        sex     cp          trtbps           chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
##  restecg    thalachh     thall   output  age_group
##  0:147   Min.   : 71.0   0:  2   0:138   A: 19
##  1:152   1st Qu.:133.5   1: 18   1:165   B:205
##  2:  4   Median :153.0   2:166           C: 79
##          Mean   :149.6   3:117
##          3rd Qu.:166.0
##          Max.   :202.0
```

# 3. Limpieza de los datos

## 3.1 Ceros y elementos vacíos

Comprobamos si en nuestro dataset tenemos valores nulos.

```
colSums(is.na(heart))
```

```
##      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
```

Podemos afirmar que no tenemos valores nulos en el dataset, por lo que no habrá que gestionarlos.

En cuanto a los valores que son 0, tenemos variables que pueden tomar este valor, como son *sex*, *cp*, *fbs*, *rest_ecg*, y *output*.

Sin embargo, la variable *thall* tiene dos registros con valor 0, lo cual no es posible según la descripción de la variable facilitada en la fuente del dataset. Lo que haremos con estos registros será apartarlos para que no sean objeto de análisis.

```
data = subset(data, thall != 0)
summary(data$thall)
```

```
##   0   1   2   3
##   0  18 166 117
```

## 3.2 Valores extremos

Cargamos la librería tidyverse, que incluye ggplot2, para poder realizar unos gráficos boxplot y observar si existen valores extremos.

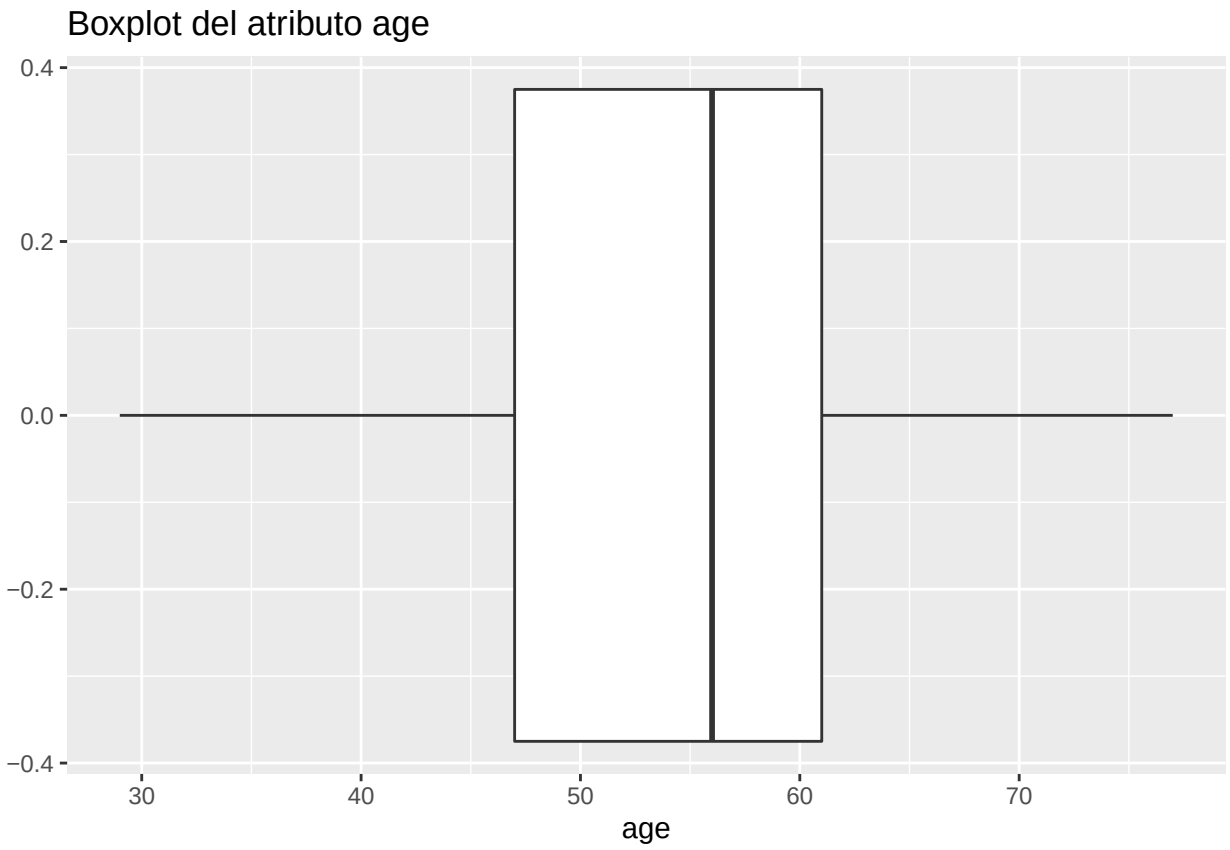
```
if(!require("tidyverse")) install.packages("tidyverse"); library("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Graficaremos las 4 variables numéricas y mostraremos los valores outlier mediante *boxplot.stats*.

```
age_box = ggplot(data, aes(x=age)) +
  geom_boxplot(outlier.colour = "red") +
  labs(title = "Boxplot del atributo age")
age_box
```

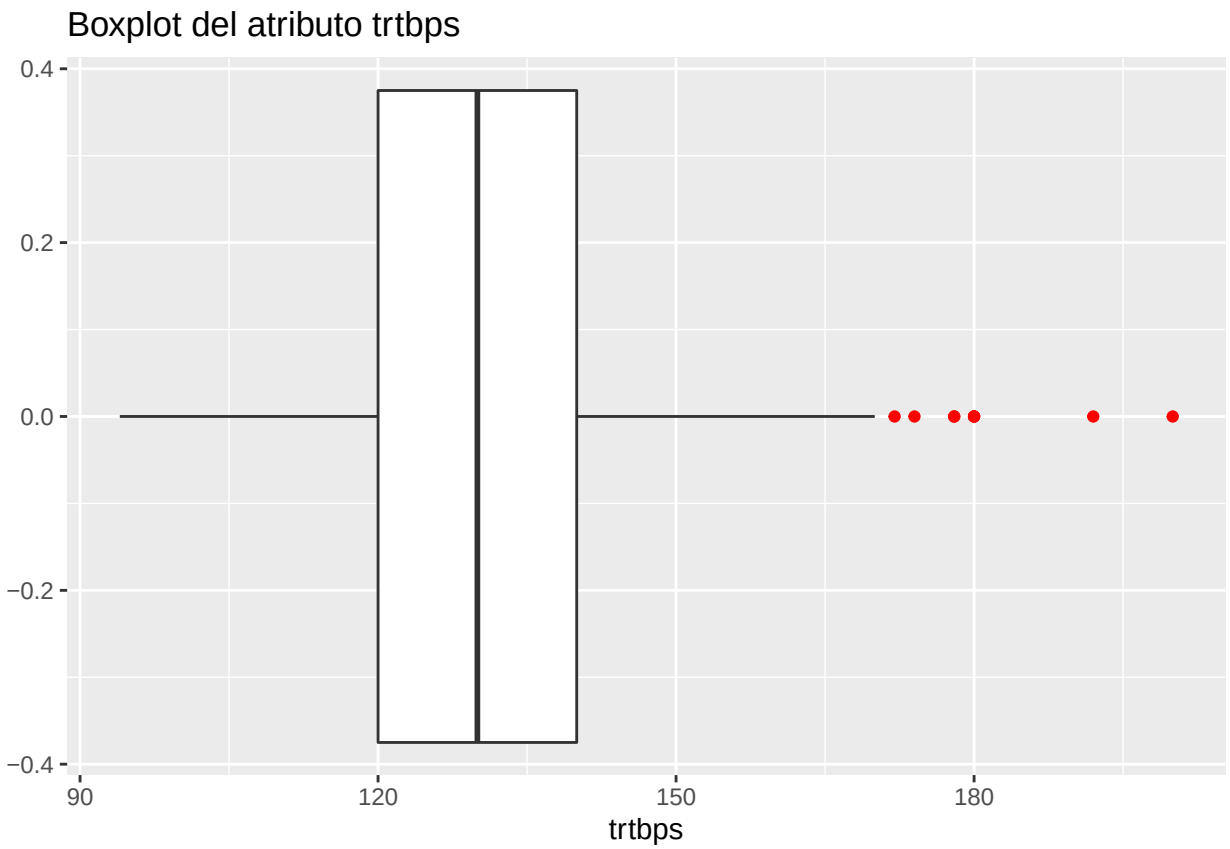



```
boxplot.stats(data$age)$out
```

```
## numeric(0)
```

En el caso de la variable *age* no tenemos valores extremos.

```
trtbps_box = ggplot(data, aes(x=trtbps)) +
  geom_boxplot(outlier.colour = "red") +
  labs(title = "Boxplot del atributo trtbps")
trtbps_box
```

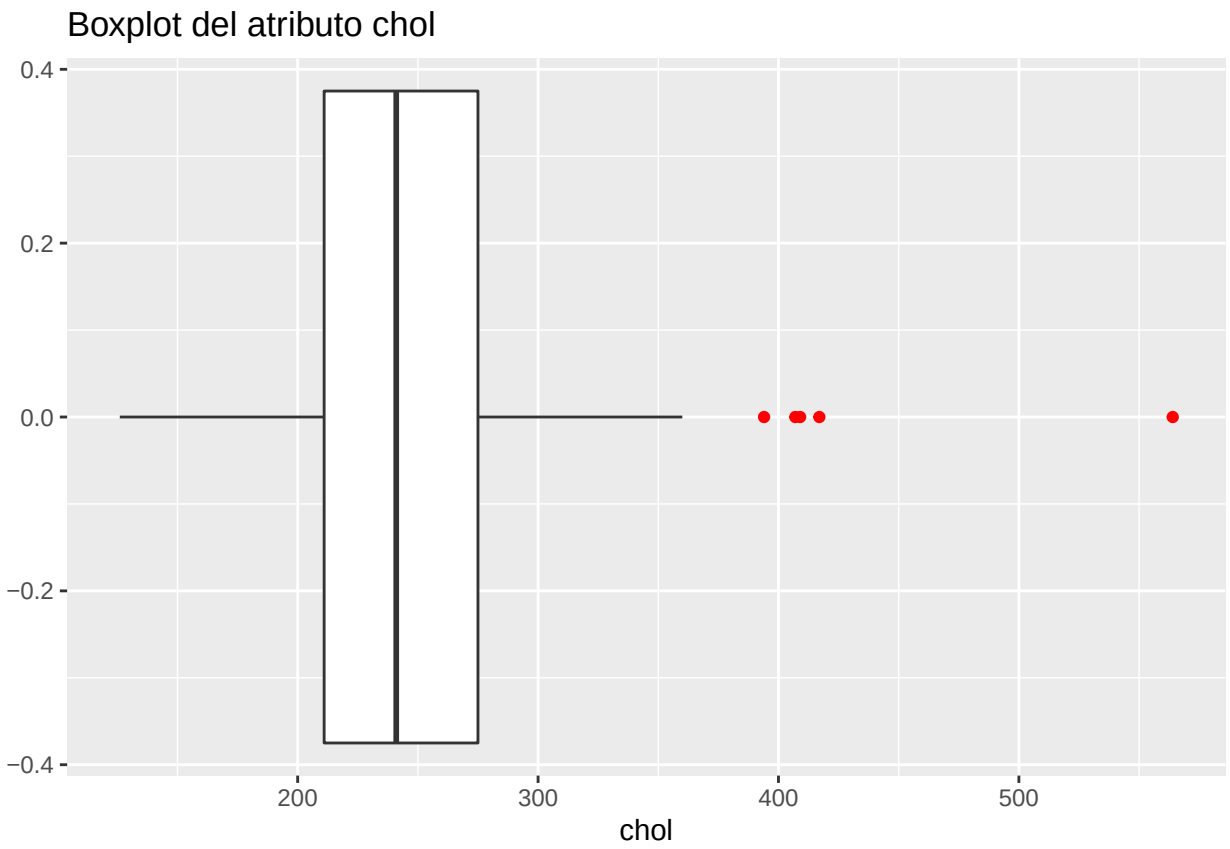



```
boxplot.stats(data$trtbps)$out
```

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

En el caso de la variable *trtbps* nos encontramos con 9 valores extremos. Aun así, parecer ser que no se tratan de errores de medición sino que únicamente son valores extremos. Los dejaremos en el dataset.

```
chol_box = ggplot(data, aes(x=chol)) +
  geom_boxplot(outlier.colour = "red") +
  labs(title = "Boxplot del atributo chol")
chol_box
```




```
boxplot.stats(data$chol)$out
```

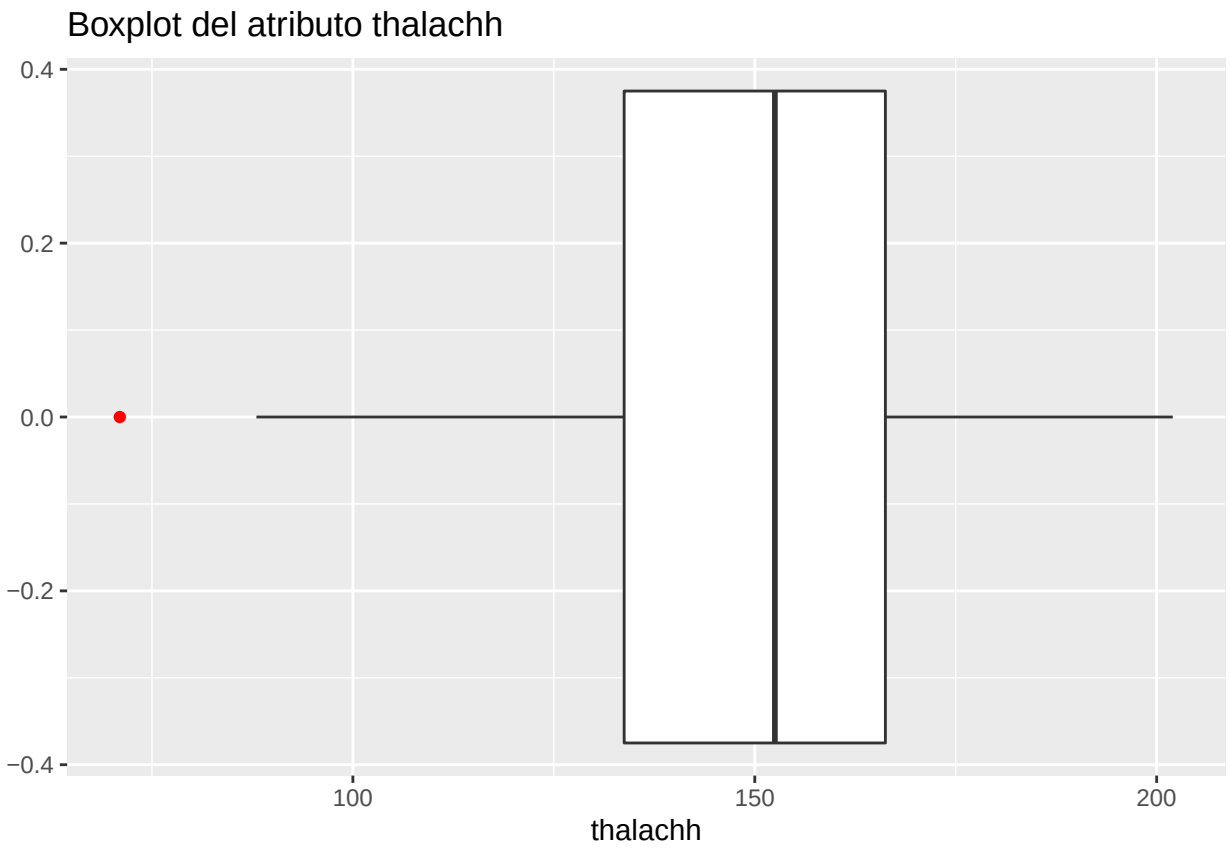
```
## [1] 417 564 394 407 409
```

En el caso de la variable *chol* tenemos 5 valores extremos, 4 de ellos son valores elevados pero muy cercanos a lo que sería un valor normal. El valor de *564* sí que puede ser extremo y no un error de medición, pero al encontrarse tan apartado y ser un único valor, lo dejaremos fuera del dataset y no será objeto de análisis. Como es un único valor, con un simple *subset* lo dejaremos fuera del dataset.

```
data = subset(data, chol != 564)
summary(data$chol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   211.0   240.5   245.4   274.2   417.0
```

```
thalachh_box = ggplot(data, aes(x=thalachh)) +
  geom_boxplot(outlier.colour = "red") +
  labs(title = "Boxplot del atributo thalachh")
thalachh_box
```




```
boxplot.stats(data$thalachh)$out
```

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

En el caso de la variable *thalachh* tenemos 1 valor extremo. Como en el caso de *trtbps* podemos asumir de que no se trata de un error sino de un valor extremo correcto, dado que se trata de una persona perteneciente al grupo de mayores de 60 años es más probable que tenga un ritmo cardíaco máximo bajo.

# 4. Análisis de los datos

Queremos realizar un estudio de la incidencia en el nivel de colesterol del sexo biológico de las personas y del rango de edad al que pertenecen.

También estamos interesados en saber si existe correlación entre alguna de las variables del conjunto de datos.

Por último, queremos encontrar un modelo de regresión que nos permita predecir la probabilidad de que una pesona sufra problemas cardíacos a partir de ciertos datos fácilmente medibles.

## 4.1. Selección de los grupos de datos

Realizamos una agrupación por sexo, distinguiendo hombres y mujeres.

```
data.male <- data[data$sex == "1",]
data.female <- data[data$sex == "0",]
```

Y realizamos también una agrupación por grupo de edad, la variable categórica que hemos creado a partir de la variable numérica *age*.

```
data.ageA <- data[data$age_group == "A",]
data.ageB <- data[data$age_group == "B",]
data.ageC <- data[data$age_group == "C",]
```

## 4.2 Normalidad y homogeneidad de la varianza

Para poder seleccionar correctamente el tipo de análisis que vamos a aplicar a los datos, es necesario comprobar si estos siguen los supuestos de distribución normal y homegeneidad de las varianzas (homocedasticidad).

Para verificar la suposición de la normalidad, algunas de las pruebas más habituales son los tests de Kolmogorov-Smirnov y de Shapiro-Wilk.

El test de Shapiro-Wilk se considera uno de los métodos más potentes para contrastar la normalidad. Asumiendo como hipótesis nula que la población está distribuida normalmente, si el p-valor es menor al nivel de significancia, generalmente $\alpha = 0.05$, entonces la hipótesis nula es rechazada y se concluye que los datos no cuentan con una distribución normal.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data$age
## W = 0.986, p-value = 0.005166
```

Como p-value $< \alpha$, se rechaza la hipótesis nula, por lo que la variable *age* no sigue una distribución normal.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data$chol
## W = 0.98343, p-value = 0.001534
```

Se obtiene la misma conclusión para la variable *chol*.

Aún así, por el Teorema Central del Límite, podemos asumir la normalidad ya que el número de observaciones es grande.

Comprobamos ahora la homocedasticidad. Al asumir la normalidad por el TCL, utilizaremos el test de Levene.

En este test, la hipótesis nula asume igualdad de varianzas en los diferentes grupos de datos, por lo que p-valores inferiores al nivel de significancia indicarán heterocedasticidad.

```
library(car)
```

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

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

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

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
leveneTest(chol ~ sex, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1   9.106 0.002768 **
##       298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que obtenemos un p-valor inferior al nivel de significancia (0.05), rechazamos la hipótesis nula de homocedasticidad y concluimos que hay heterogeneidad de varianzas.

## 4.3 Análisis estadístico comparativo

### 4.3.1 ¿Tienen los hombres el colesterol más alto que las mujeres?

En este caso realizaremos una comparación entre dos grupos.

Como hemos comprobado que no se cumple el criterio de homocedasticidad, aplicaremos una prueba no paramétrica como es el test de Mann-Whitney, ya que se trata de grupos de datos independientes.

```
wilcox.test(chol ~ sex, data = data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  chol by sex
## W = 11454, p-value = 0.01105
## alternative hypothesis: true location shift is not equal to 0
```

El valor de p-value menor que el nivel de significancia indica que hay que rechazar la hipótesis nula de que el colesterol es igual para los dos grupos, y por tanto podemos concluir que hay diferencias entre los dos sexos.

### 4.3.2 ¿Hay diferencias en el colesterol entre los diferentes grupos de edad?

En este caso vamos a aplicar un análisis de comparación entre más de dos grupos.

Al no cumplirse las asunciones de normalidad e igualdad de varianzas, hemos de utilizar un test no paramétrico, por lo que utilizaremos el test de Kruskal-Wallis.

```
kruskal.test(chol ~ age_group, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  chol by age_group
## Kruskal-Wallis chi-squared = 14.043, df = 2, p-value = 0.0008927
```

Dado que el p-valor obtenido es menor al nivel de significancia, se puede concluir que el nivel de colesterol muestra diferencias significativas para los diferentes grupos de edad.

### 4.3.3 Correlación

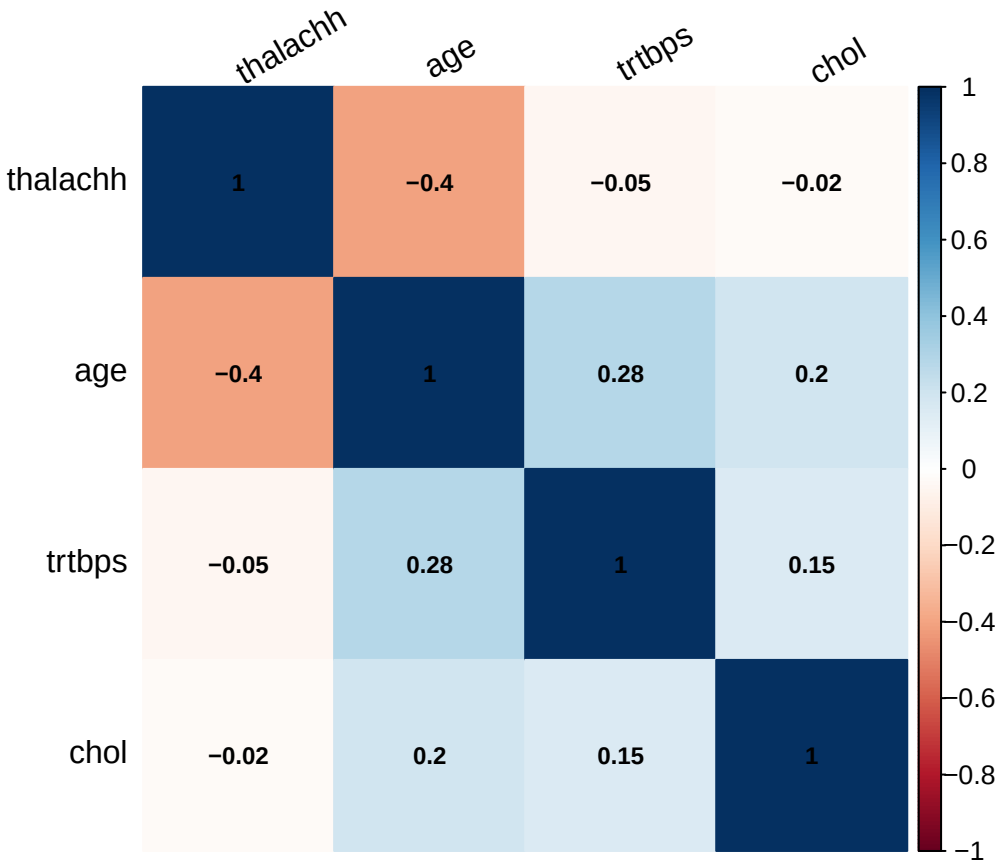
Elaboramos un gráfico de correlación, para encontrar posibles correlaciones entre variables numéricas. Para ello, cargamos la librería *corrplot*.

```
if(!require("corrplot")) install.packages("corrplot"); library("corrplot")
```

```
## Loading required package: corrplot
```

```
## corrplot 0.92 loaded
```

```
n = c("age", "chol", "trtbps", "thalachh")
factores = data %>% select(all_of(n))
res = cor(factores)
corrplot(res,method="color",tl.col="black", tl.srt=30, order = "AOE", number.cex=0.75, sig.level = 0.01
```




La correlación más fuerte que encontramos es de tipo negativa entre las variables *age* y *thalachh*. Esta observación tiene sentido, ya que, tal y como hemos comentado anteriormente, es normal que a mayor edad

menor sea el ritmo cardíaco máximo obtenido.

#### 4.3.4 Regresión.

Mediante un modelo de regresión logística intentaremos predecir el valor de *output* en función del resto de parámetros.

Primero generamos los conjuntos de entrenamiento y de test. Con un dataset de tan pocas observaciones es complicado, pero veremos qué resultados obtenemos.

```
n<- dim(data)[1]
set.seed(1234)
train <- sample(1:n , 0.8*n)
data.test <- data[-train,]
data.train <- data[train,]
```

Elaboramos un primer modelo de regresión logística con todas las variables.

```
model_log <- glm(output ~ . -age_group, family = binomial(link = logit), data= data.train)
summary(model_log)
```

```
##
## Call:
## glm(formula = output ~ . - age_group, family = binomial(link = logit),
##     data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1758  -0.5068   0.1898   0.5494   2.6535
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.514e-01  2.667e+00   0.057  0.95474
## age         -2.057e-02  2.525e-02  -0.815  0.41533
## sex1        -1.454e+00  5.058e-01  -2.875  0.00404 **
## cp1          2.470e+00  6.339e-01   3.897 9.74e-05 ***
## cp2          1.980e+00  4.568e-01   4.335 1.46e-05 ***
## cp3          1.609e+00  6.368e-01   2.527  0.01152 *
## trtbps      -1.534e-02  1.070e-02  -1.434  0.15163
## chol        -7.265e-03  4.416e-03  -1.645  0.09995 .
## fbs1        -1.049e-01  5.791e-01  -0.181  0.85625
## restecg1     2.127e-01  3.889e-01   0.547  0.58435
## restecg2    -1.426e+01  1.011e+03  -0.014  0.98875
## thalachh     3.300e-02  1.095e-02   3.014  0.00258 **
## thall2       7.522e-01  8.245e-01   0.912  0.36162
## thall3      -9.401e-01  7.783e-01  -1.208  0.22711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.44  on 239  degrees of freedom
## Residual deviance: 183.43  on 226  degrees of freedom
## AIC: 211.43
##
## Number of Fisher Scoring iterations: 14
```

Quitamos las variables no explicativas como *age*, *trtbps*, *chol*, *fbs*, *restecg*, *thall* y *age_group*.

```
model_log <- glm(output ~ . -age_group -age -trtbps - chol - fbs -restecg - thall, family = binomial(li
summary(model_log)
```

```
##
## Call:
## glm(formula = output ~ . - age_group - age - trtbps - chol -
##     fbs - restecg - thall, family = binomial(link = logit), data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5067  -0.7056   0.2667   0.6575   2.4612
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.907379   1.313576  -4.497 6.89e-06 ***
## sex1        -1.599715   0.394131  -4.059 4.93e-05 ***
## cp1          2.650032   0.583397   4.542 5.56e-06 ***
## cp2          2.095230   0.402219   5.209 1.90e-07 ***
## cp3          1.610210   0.569581   2.827   0.0047 **
## thalachh     0.040793   0.008845   4.612 3.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.44  on 239  degrees of freedom
## Residual deviance: 217.31  on 234  degrees of freedom
## AIC: 229.31
##
## Number of Fisher Scoring iterations: 5
```

Probamos la predicción del modelo utilizando para ello el conjunto de test.

```
data.test$pred <- predict(model_log,data.test, type="response")
data.test$pred_final <- ifelse(data.test$pred > 0.5, 1, 0)
table(data.test$pred_final,data.test$output, dnn=c("Pred.","Obs."))
```

```
##      Obs.
## Pred.  0  1
##     0 24  5
##     1  7 24
```

Obtenemos los siguientes resultados:

- Sensibilidad = VP/(VP+FN)
- VP = 30/33 = 0.91
- Predice correctamente el 91% de los casos.

## Conclusiones

Al analizar la estructura del data frame, hemos podido ver que el conjunto de datos constaba tanto de variables calitativas como cuantitativas, y que en ninguna de ellas aparecían valores ausentes, y muy pocos valores extremos que se pudiesen considerar erróneos, y que al ser un número muy pequeño, hemos decidido eliminarlos del conjunto de datos.

Separamos los datos en grupos por sexos (utilizando el factor sex) y por grupos de edad (utilizando una variable categórica age_group generada a partir de la variable numérica age), con la finalidad de realizar un estudio de variabilidad del nivel de colesterol en sangre en función del sexo y del grupo de edad al que pertenecen los indivíduos de la población.

Hemos comprobado que la distribución de las variables no cumplía con el supuesto de normalidad, aunque la podríamos asumir por el Teorema Central del Límite, también encontramos que no cumplen con el criterio de homocedasticidad, por lo que necesitaremos aplicar tests no paramétricos en los análisis.

A partir de los distintos análisis, hemos podido comprobar que sí que existe una diferencia en el nivel de colesterol entre hombres y mujeres, y que también existe diferencia entre distintos grupos de edad.

Se observa una correlación negativa entre la edad y la frecuencia cardíaca máxima.

Por último, hemos obtenido un modelo de regresión logística que nos permita predecir, a partir de una serie de valores, la probabilidad de que el sujeto sufra un problema cardíaco.

A partir de la generación del modelo de regresión, hemos visto que muchas de las variables del modelo no son significativas para predecir los problemas cardíacos, quedando como variables explicativas únicamente *sex*, *cp* y *thallach*, por lo que es probable que el conjunto de datos no sea muy útil para resolver el problema planteado.

## Exportación del fichero de datos resultante

```
write.csv(data, file="../data/heart_final.csv")
```

## Contribuciones

| Contribuciones | Firma |
|---|---|
| Investigación Previa | NCS,DMM |
| Redacción de las respuestas | NCS,DMM |
| Desarrollo del código | NCS,DMM |
| Participación en el vídeo | NCS,DMM |