Práctica 2: Limpieza y análisis de datos

Maider Dorronsoro, Flavia Felletti

2023-06-14

Contents

1 Descripción del dataset.

El dataset seleccionado son datos relativos a pacientes de diferentes países, en concreto a pacientes con alguna enfermedad cardiovascular. Según los datos de la World Health Organization (WHO), las enfermedades cardiovasculares son la principal causa de muerte en el mundo. Se ha calculado que cada año al rededor de 17.9 millones de personas pierden su vida por alguna enfermedad cardiovascular.Además, un tercio de ellos, son personas menores a 70 años, lo cual provoca una muerte temprana.Por ello, creemos importante e interesante analizar esta problemática.

Al fin y al cabo, en este proyecto se va a analizar/detectar los factores que más aumentan la probabilidad de padecer de enfermedades cardiovasculares El dataset ha sido obtenido de esta fuente: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

```
df<-read.csv("heart.csv")
df$sex <- as.character(df$sex)
```

A continuación, se procede a hacer un pequeño análisis del dataset:

```
# visualizo las dimensiones del dataset:
dimdat <- dim(df)
dimdat
```

```
## [1] 303  14
```

Tenemos un conjunto de datos con 303 registros y 14 variables a analizar, de las cuales solo vamos a describir las que nos interesan en este análisis.

- Age:Edad de los pacientes
- Sex:Género pacientes [0: F, 1: M]
- trtbps: Resting blood pressure- Presión arterial en reposo [mm Hg]
- chol: Colesterol [mm/dl]
- restecg: Resultados del electrocardiograma en reposo [0: Normal, 1: con anormalidad de la onda ST-T, 2: muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes].
- thalachh: Frecuencia cardíaca máxima alcanzada [Valor numérico entre 71 y 202]
- output: Enfermedad cardiaca [1: tiene enfermedad cardíaca, 0: no tiene enfermedad cardíaca]

A continuación un resumen estadístico de las diferentes variables a analizar:

```
summary(df)
```

```
##       age            sex                 cp            trtbps
##  Min.   :29.00   Length:303         Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   Class :character   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Mode  :character   Median :1.000   Median :130.0
##  Mean   :54.37                      Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00                      3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00                      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
```

---

2 Integración y selección de los datos de interés a analizar.

En este proyecto se quiere analizar la realidad de las enfermedades cardiovasculares de las personas en relación a las variables seleccionadas.Como unicamente nos interesan las mencioandas, se hace una subselección de estas.

```
df <- select (df,'age','sex', 'chol','restecg','trtbps','thalachh','output')
df$sex <- ifelse(df$sex=='0','F','M')
head(df)
```

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

---

3 Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

A continuación se procede a analizar la existencia de valores perdidos NA, NULL o blancos:

```r
# averiguo si el dataset contiene valores NA ("not available")
any(is.na(df))
```

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

```r
# averiguo si hay valores NULL
any(is.null(df))
```

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

```r
# averiguo si hay valores blancos
any(df == "")
```

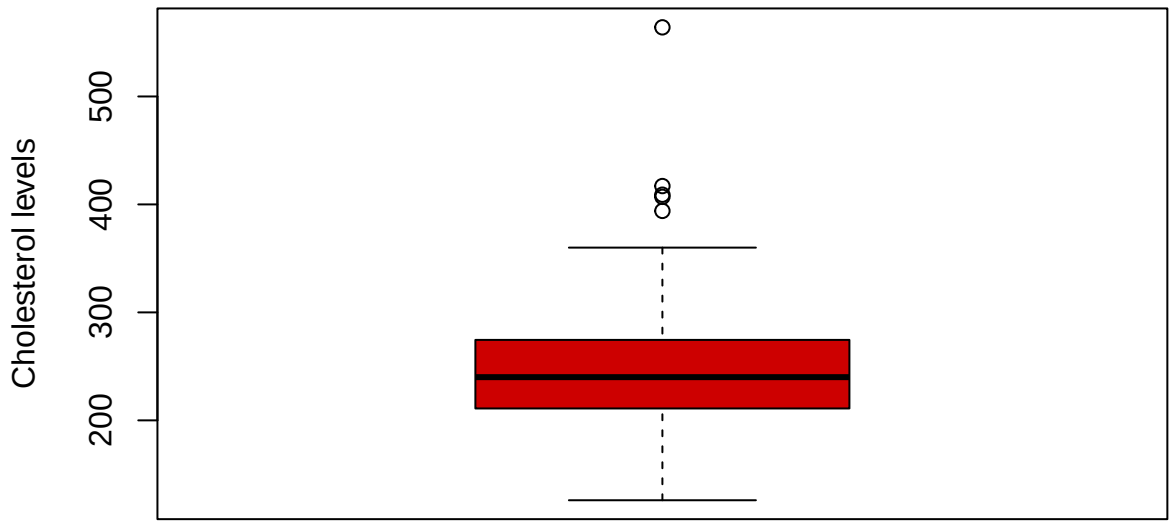
```
## [1] FALSE
```

Tal y como se puede observar no hay registros incompletos, por lo que no se van a tratar.

3.2. Identifica y gestiona los valores extremos

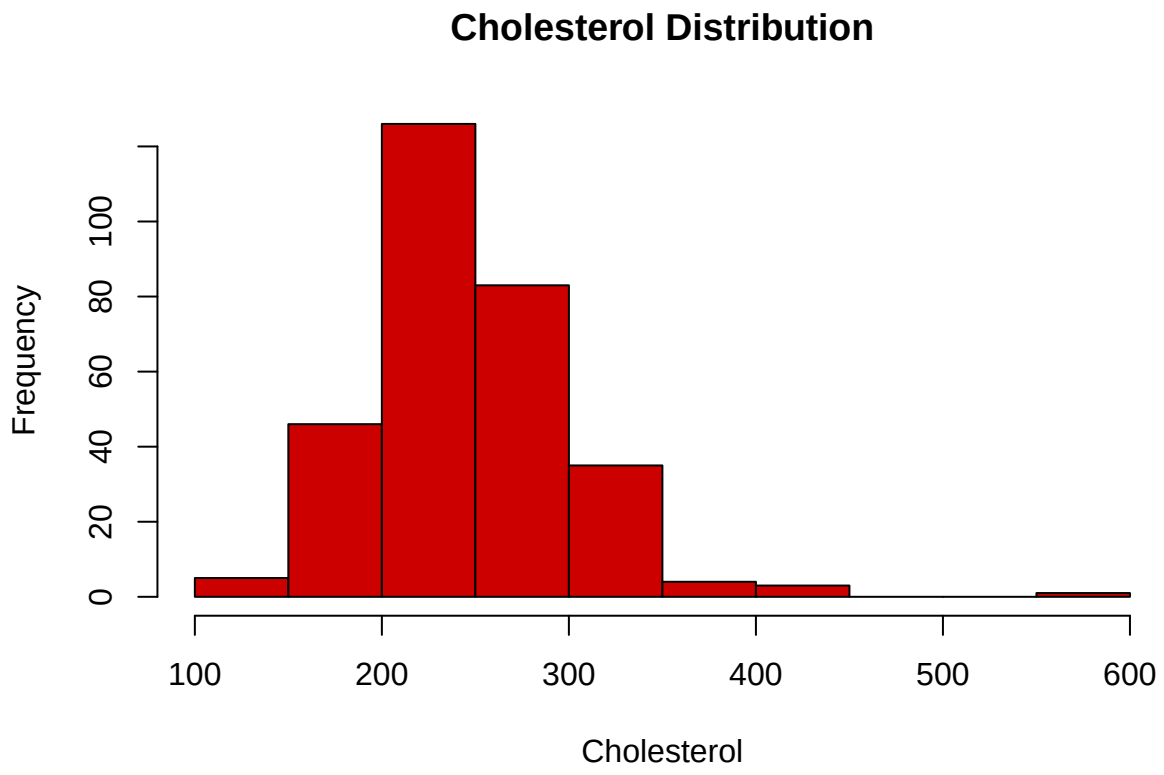
Diagrama de cajas de la variable Cholesterol

```r
boxplot(df$chol,
  ylab = "Cholesterol levels",
  col = "red3"
)
```



Histograma de la variable Cholesterol

```
hist(df$chol, xlab="Cholesterol",
     main="Cholesterol Distribution", col="red3")
```



Se puede observar que existen valores de colesteról que superan los 400 mm/dl. Esto no tiene porqué ser resultado de un error; de hecho, el colesteról tan alto podría ser causa de una enfermedad cardiaca grave.[1]

Otros casos de incosistencisa o datos incoherentes podrían ser los valores de restecg (presión arterial en en reposo) y Chol (colesterol) iguales a 0.

A continuación se analiza la cantidad de resgistros que tienen colesterol igual a 0 y/o presión arterial en reposo igual a 0.

```
# Cholesterol equal zero
"Cholesterol:"
```

```
## [1] "Cholesterol:"
```

```
sum(df$chol == 0)
```

```
## [1] 0
```

[1] según este artículo en la sección "health" de CNN, hay condiciones genéticas que pueden provocar niveles de hasta 600 mg/dL http://edition.cnn.com/2009/HEALTH/11/24/moh.healthmag.cholesterol.surprises/index.html#:~:text=But%20for%20some%20families%2C%20it's,heart%20attacks%20early%20in%20life.

```r
# RestingBP equal zero
"RestingBP:"
```

```
## [1] "RestingBP:"
```

```r
sum(df$trtbps == 0)
```

```
## [1] 0
```

Contamos el número de observaciones que tienen colesteról inferior a 40 mg/dL, que equivale a niveles muy bajos de colesteról, aunque probables.[2]

```r
sum(df$chol < 40)
```

```
## [1] 0
```

Se eliminarán del conjunto de datos las observaciones que tienen niveles de colesterol o de presión cardíaca en reposo iguales a cero y se vuelve a visualizar el resúmen estadistico.

```r
# creates a new dataset deleting Cholesterol = 0 and RestingBP = 0 from
# the original dataset
dff <- df[(df$chol != 0 & df$restecg != 0),]

# visualizes summary
summary(dff)
```

```
##       age            sex                 chol          restecg
##  Min.   :34.00   Length:156         Min.   :126.0   Min.   :1.000
##  1st Qu.:45.00   Class :character   1st Qu.:204.0   1st Qu.:1.000
##  Median :54.00   Mode  :character   Median :232.0   Median :1.000
##  Mean   :53.12                      Mean   :237.9   Mean   :1.026
##  3rd Qu.:60.00                      3rd Qu.:266.2   3rd Qu.:1.000
##  Max.   :76.00                      Max.   :354.0   Max.   :2.000
##      trtbps         thalachh         output
##  Min.   : 94.0   Min.   : 71.0   Min.   :0.0000
##  1st Qu.:120.0   1st Qu.:138.8   1st Qu.:0.0000
##  Median :128.5   Median :157.5   Median :1.0000
##  Mean   :129.4   Mean   :151.3   Mean   :0.6218
##  3rd Qu.:140.0   3rd Qu.:169.0   3rd Qu.:1.0000
##  Max.   :180.0   Max.   :194.0   Max.   :1.0000
```

Visualización de las variables

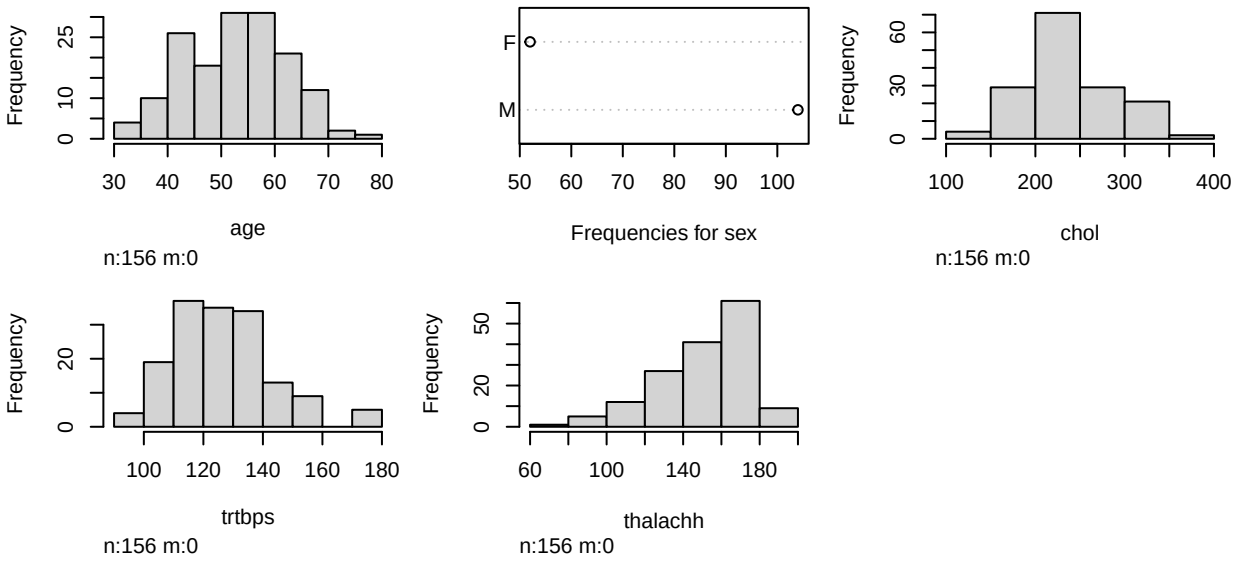
Visualizamos una premera representación gráfica de la distribución de los datos por cada variable:

```
##
## Attaching package: 'Hmisc'
```

[2]Según el artículo: https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/expert-answers/cholesterol-level/faq-20057952

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```



---

4 Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar

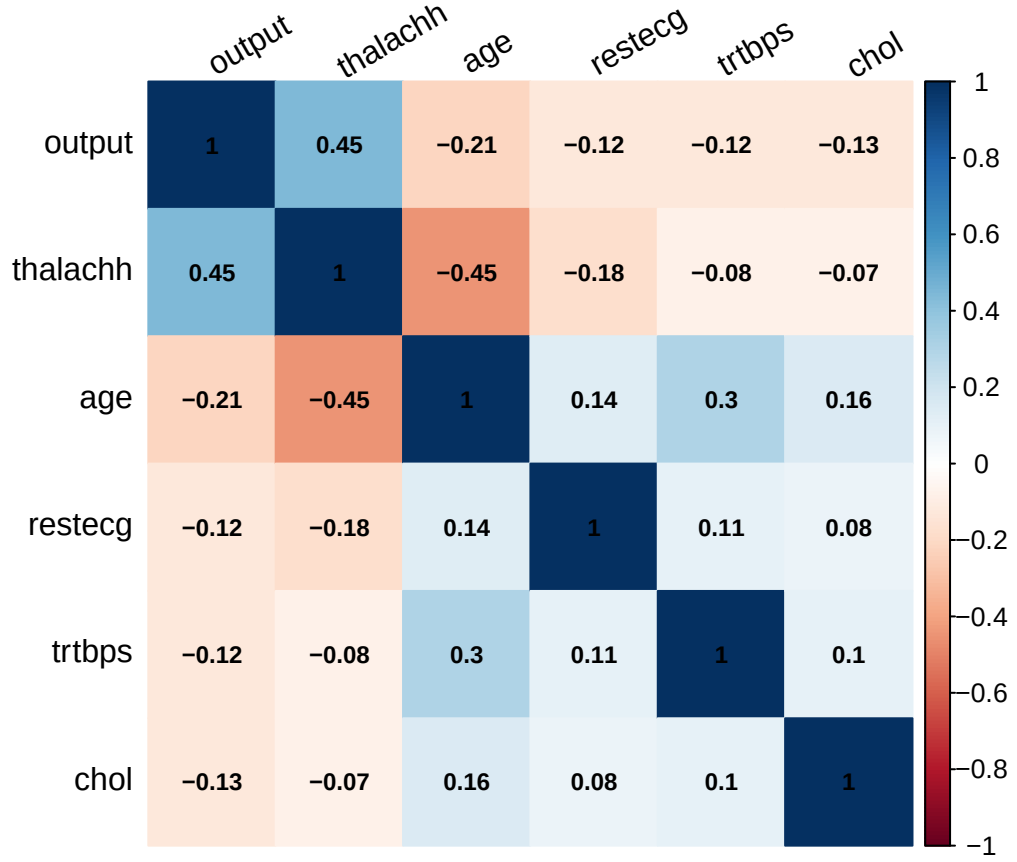
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

Los datos que se van a utilizar en los siguientes análisis es el dataset ya limpio de los anteriores apartados. Además, para cada análisis que se modificará como sea necesario el dataframe.

Análisis de correlaciones

```
## corrplot 0.92 loaded
```



Se puede observar que hay una correlación moderada entre la variable output (enfermedad cardíaca) y la variabla thalachh (frecuencia cardíaca máxima alcanzada), ya que está llega a 0,45.

Análisis de regresión logística

Una vez analizadas las correlaciones vamos a calcular la regresión logística para calcular la variable de salida, la cual es dicotomica:

```
modelo <- glm(dff$output ~ dff$age + dff$sex + dff$chol +
                dff$restecg + dff$trtbps + dff$thalachh, data=dff)
summary(modelo)
```

```
##
## Call:
## glm(formula = dff$output ~ dff$age + dff$sex + dff$chol + dff$restecg +
##     dff$trtbps + dff$thalachh, data = dff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9764  -0.3459   0.1218   0.3206   0.8944
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4078846  0.5267400   0.774 0.439947
## dff$age      -0.0011712  0.0043163  -0.271 0.786509
## dff$sexM     -0.2817279  0.0732739  -3.845 0.000178 ***
## dff$chol     -0.0010729  0.0007203  -1.490 0.138459
## dff$restecg  -0.2143403  0.2191848  -0.978 0.329710
## dff$trtbps   -0.0026384  0.0021204  -1.244 0.215335
## dff$thalachh  0.0084627  0.0016328   5.183    7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1757739)
##
##     Null deviance: 36.686  on 155  degrees of freedom
## Residual deviance: 26.190  on 149  degrees of freedom
## AIC: 180.33
##
## Number of Fisher Scoring iterations: 2
```

Tal y como se puede observar las variables más relevantes son el sexo y el thalachh, ya que son las únicas que han resultado significativas.

$$(P_value < 0.05)$$

Además, se cumple la estimación que se había hecho en estudio de correlaciones, ya que la variable thalachh tiene un impacto positivo, de forma que cuanto mayor sea, mayor es la probabilidad de padecer una enfermedad cardiovascular.

Análisis de contraste de hipótesis de dos poblaciones

Por último, se va a anlizar analizar si la edad media de los pacientes es la misma independientemente del sexo. Es decir, se analizará si la media de edad de los hombres y mujeres enfermos es la misma (output=1):

```
## 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   8.097 0.005431 **
##       95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
##  Shapiro-Wilk normality test
##
## data:  dff_sick$age
## W = 0.97444, p-value = 0.055
```

Gracias a estas pruebas se ha visto que aunque la varianza de esta variable sea homogenea, es decir, y sigue una distribución normal. Ello implica que podemos utilizar el contraste de hipótesis de medias usual:

$$H_1 : \mu_m \neq \mu_f H_0 : \mu_m = \mu_f$$

En concreto, la prueba que se va a realizar es un contraste de hipótesis bilateral:

```
#Generamos dos grupos
dff1<- subset(dff,dff$output=='1')
df_m <- subset(dff1,dff1$sex == 'M')
df_f <- subset(dff1,dff1$sex == 'F')

#COmo no sigue una normal, hacemos uso del test Wilcox,
#que analiza si las medianas de estos dos grupos son diferentes.
t.test(df_m$age, df_f$age)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_m$age and df_f$age
## t = -2.0762, df = 67.929, p-value = 0.04166
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.3476793 -0.1653869
## sample estimates:
## mean of x mean of y
##  49.76786  54.02439
```

Como el

$$p - value$$

sale menor a 0.05, se podría aceptar que hay diferencia de edad media en entre las hombres enfermos con una enfermedad cardiovascular y las mujeres.

5. Representación de los resultados a partir de tablas y gráficas.

A lo largo de toda la práctica se han utilizado visualizaciones.

6. Resolución del problema.

El problema presentado en un inicio era el análisis de los pacientes enfermos de una enfermedad cardiovascular, a continuación se especifican los resultados de dicho análisis:

Para empezar, se ha visto que la variable que más efecto tiene sobre la variable dependiente, es thalachh, la frecuencia cardíaca máxima alcanzada de las personas. Esto se ha podido ver tanto en el análisis de correlaciones como en el modelo logístico.

Por otro lado, gracias al análisis de contrastes de media realizado se ha podido concluir que la edad media de los hombres que padecen enfermedades cardiovasculares es igual al de las mujeres, es decir, no hay una significancia estadística.