Universitat Oberta de Catalunya

uoc.edu

Tipología y ciclo de vida de los datos

Práctica 2: Integración, Limpieza y Análisis del dataset "Heart Attack Analysis and Prediction Dataset"

Autores:

Daniel Romero Resano

Fabrizio Jesus Cáceda Peña

1. Descripción del dataset. ¿Por qué es importante y qué pregunta / problema pretende responder?

El conjunto de datos que se va a utilizar para el análisis es “Heart Attack Analysis & Prediction Dataset”, este dataset está formado por 14 variables los cuales muestran la clasificación de ataques al corazón.
Actualmente las muertes cardiovasculares son una de las principales muertes en todo el mundo. Conociendo este hecho, con el conjunto de datos que disponemos pretendemos predecir, gracias a las distintas características que contiene, si el paciente va a padecer un ataque al corazón o no.

En este conjunto encontramos datos de tipo numérico, de tipo binario y de tipo categórico:

- Age : Edad del paciente. (Cuantitativa)
- Sex : Género del paciente. (Cualitativa)
- exng: angina inducida por el ejercicio (1 = sí; 0 = no). (Cualitativa)
- caa: número de buques principales (0-4). (Cualitativa)
- cp : Tipo de dolor torácico tipo de dolor torácico. (Cualitativa)
    - Valor 1: angina típica
    - Valor 2: angina atípica
    - Valor 3: dolor no anginoso
    - Valor 4: asintomático
- trtbps : presión arterial en reposo (en mm Hg). (Cuantitativa)
- chol : colesterol en mg/dl obtenido a través del sensor BMI. (Cuantitativa)
- fbs : (azúcar en sangre en ayunas > 120 mg/dl) (1 = verdadero; 0 = falso). (Cualitativa)
- oldpeak: Pico previo. (Cuantitativa)
- rest_ecg : resultados electrocardiográficos en reposo. (Cualitativa)
    - Value 0: normal
    - Value 1: tener anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0,05 mV)
    - Value 2: mostrar hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes
- thalach : frecuencia cardíaca máxima alcanzada. (Cuantitativa)
- target : 0 = menos posibilidades de ataque al corazón 1= más posibilidades de ataque al corazón. (Cualitativa)
- slp: Slope.
- thall: Thal rate.

```
> dim(dataHeart)
[1] 303  14
```

Viendo la dimensión del dataset, el conjunto está formado por 303 observaciones, cada una representa un paciente, y 14 características.

Como podemos ver a continuación, las variables del dataset no están codificadas adecuadamente de manera que las tendremos que adecuar al tipo de datos que contiene.

```
> str(dataHeart)
'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 ...
```

A continuación se codifican cada una de las variables según el tipo de dato que contienen, de manera que los datos quedan así:

```
> str(dataHeart)
'data.frame':   303 obs. of  14 variables:
 $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : logi  TRUE TRUE FALSE TRUE FALSE TRUE ...
 $ cp      : Factor w/ 4 levels "0","1","2","3": 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 : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
 $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
 $ exng    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
 $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slp     : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
 $ caa     : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ thall   : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ output  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

2. Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

Del conjunto de datos que disponemos, hemos decidido no contar con las variables de las tomas de oxígeno (o2Saturation.csv) y analizar entonces los datos sobre ataques al corazón respecto a las variables del archivo (heart.csv).

3. Limpieza de los datos.

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

Primero realizaremos la lectura del fichero y carga de datos:

```
dataHeart<-read.csv("heart.csv",header=T,sep=",")
attach(dataHeart)
```

Pasamos a probar si hay valores NA en los datos:

```
> sapply(dataHeart, function(x) sum(is.na(x)))
     age      sex       cp   trtbps     chol      fbs  restecg thalachh     exng  oldpeak      slp      caa    thall   output
       0        0        0        0        0        0        0        0        0        0        0        0        0        0
```

Podemos observar que no hay ningún NA en las 14 dimensiones con las que cuentan los datos.

Ahora comprobamos si existen valores vacíos:

```
sapply(dataHeart, function(x) sum(x==""))
     age      sex       cp   trtbps     chol      fbs  restecg thalachh     exng  oldpeak      slp      caa    thall   output
       0        0        0        0        0        0        0        0        0        0        0        0        0        0
```

Podemos observar que no existen campos vacíos.

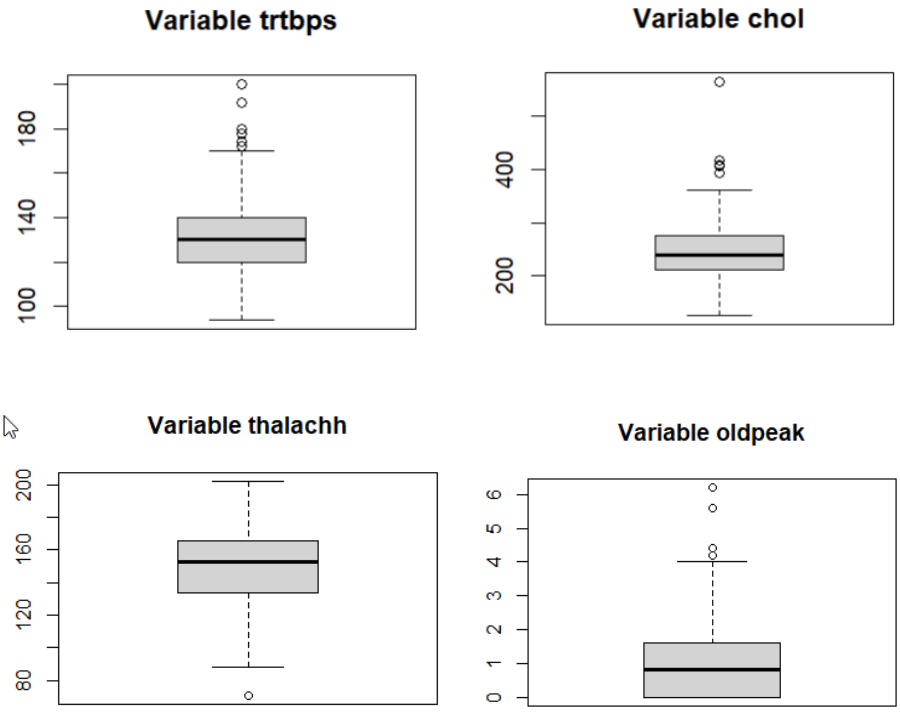
3.2. Identifica y gestiona los valores extremos.

A continuación, comprobaremos si entre nuestras características existen valores extremos o *outliers,* que son esas observaciones que se encuentran muy alejadas de la media de los datos, o que no tienen sentido con la resta.

Comprobaremos para cada característica numérica si existen estos valores:

```
> boxplot.stats(dataHeart$age)$out
integer(0)
```

```
> boxplot.stats(dataHeart$trtbps)$out
[1] 172 178 180 180 200 174 192 178 180
> boxplot(x = dataHeart$trtbps)
```




```
> boxplot.stats(dataHeart$chol)$out
[1] 417 564 394 407 409
> boxplot(x = dataHeart$chol)
```

```
> boxplot.stats(dataHeart$thalachh)$out
[1] 71
```

```
> boxplot.stats(dataHeart$oldpeak)$out
[1] 4.2 6.2 5.6 4.2 4.4
```

```
boxplot(x = dataHeart$chol,main='Variable chol')
out_hsize <- boxplot.stats(dataHeart$chol)$out
out_ind_hsize <- which(dataHeart$chol %in% c(out_hsize))

x<-dataHeart
x<- x[-which(x$chol %in% out_hsize),]
dataHeart <- x
```

Según la variable trtbps (presión arterial en reposo), podemos observar valores que podrían ser considerados como outliers por métodos estadísticos.

Sin embargo, estos valores pueden considerarse como casos de hipertensión de grado 3 y estar muy relacionados con ataques al corazón. Por ello, se mantendrán estos valores altos en la muestra de estudio.

Por otro lado, la edad está en el rango entre los 27 y 77 años y no se observa ningún valor atípico, por lo tanto no hay nada a modificar para esta variable.

La variable chol, muestra los resultados de colesterol en mg/dl obtenido a través del sensor BMI, hemos visto que hay unos pocos valores atípicos que se encuentran por encima de la media de manera que hemos decidido eliminarlos.

En cuanto a la thalachh (Frecuencia cardiaca máxima), aunque atípicos estos valores bajos pueden ser también normales en deportistas y tener sentido en la muestra, por tanto no se modificarán.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (p.ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)

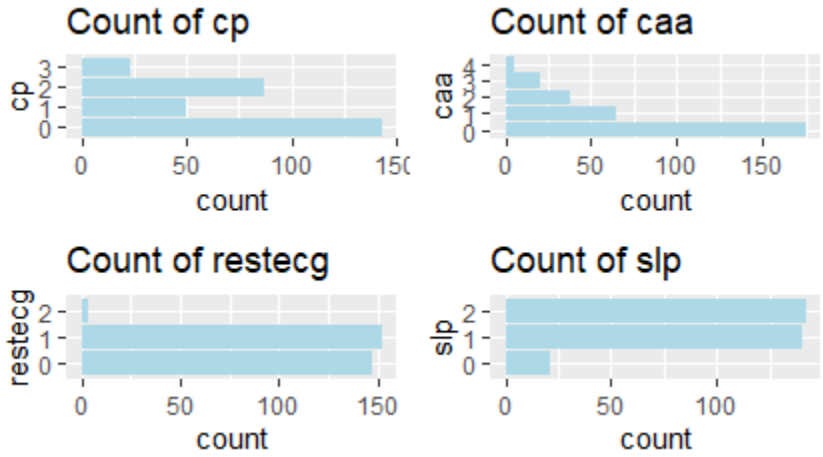
Empezamos por realizar un análisis univariable de cada uno de los atributos del dataset:

```
> summary(dataHeart)
      age             sex               cp             trtbps           chol            fbs              restecg      
 Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0   1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :55.00   Median :1.0000   Median :1.000   Median :130.0   Median :240.0   Median :0.0000   Median :1.0000  
 Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6   Mean   :246.3   Mean   :0.1485   Mean   :0.5281  
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0   3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0   Max.   :564.0   Max.   :1.0000   Max.   :2.0000  
    thalachh         exng           oldpeak          slp             caa             thall           output     
 Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000  
 1st Qu.:133.5   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000  
 Median :153.0   Median :0.0000   Median :0.80   Median :1.000   Median :0.0000   Median :2.000   Median :1.0000  
 Mean   :149.6   Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294   Mean   :2.314   Mean   :0.5446  
 3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.0000  
 Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000   Max.   :3.000   Max.   :1.0000  
```

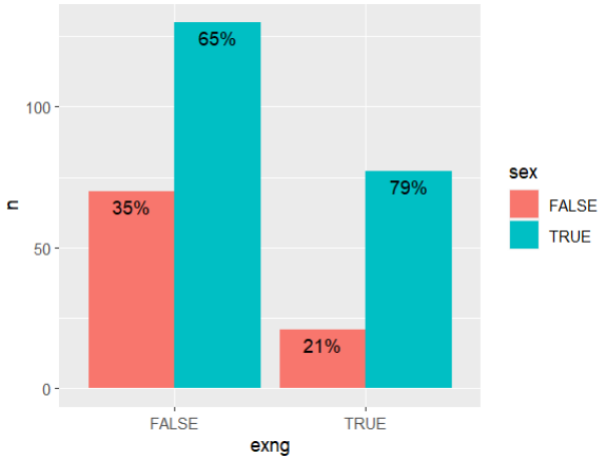
```
> table(dataHeart$exng)

FALSE  TRUE
  204    99
> table(dataHeart$fbs)

FALSE  TRUE
  258    45
> table(dataHeart$sex)

FALSE  TRUE
   96   207
```

```
> table(dataHeart$output)

FALSE  TRUE
  138   165
> table(dataHeart$thall)

FALSE  TRUE
    2   301
> table(dataHeart$caa)

  0   1   2   3   4
175  65  38  20   5
```

```
> table(dataHeart$cp)

  0   1   2   3
143  50  87  23
> table(dataHeart$restecg)

  0   1   2
147 152   4
> table(dataHeart$slp)

  0   1   2
 21 140 142
```

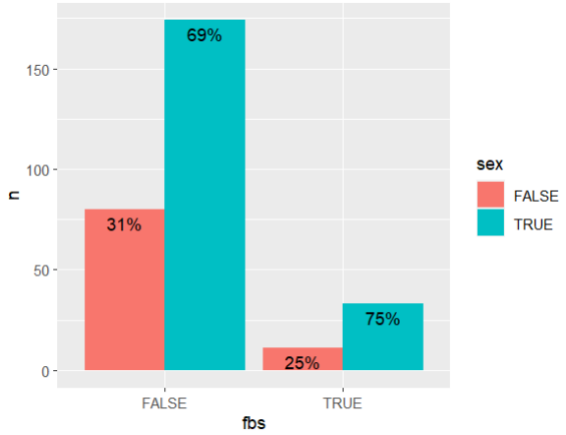



```
test_grafica<-dataHeart %>%
  group_by(exng) %>%
  count(sex)%>%
  mutate(porcentaje=scales::percent(n/sum(n)))

ggplot(test_grafica,aes(x=exng , y=n, fill=sex))+
  geom_bar(stat="identity", position="dodge")+
  geom_text(aes(label=porcentaje),color="black", vjust=1.5, position = position_dodge(0.9))
```



Como se puede ver en la gráfica 4 de cada 5 pacientes de género sex = 1, padecen anginas inducidas por la realización de ejercicio, mientras que 2 de cada 3 pacientes de género sex = 1, no lo padecen.

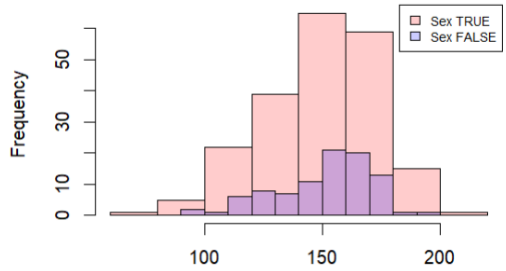
```
test_grafica<-dataHeart %>% group_by(fbs) %>% count(sex)%>%
  mutate(porcentaje=scales::percent(n/sum(n)))

ggplot(test_grafica,aes(x=fbs , y=n, fill=sex))+ geom_bar(stat="identity", position="dodge")+
geom_text(aes(label=porcentaje),color="black", vjust=1.5, position = position_dodge(0.9))
```



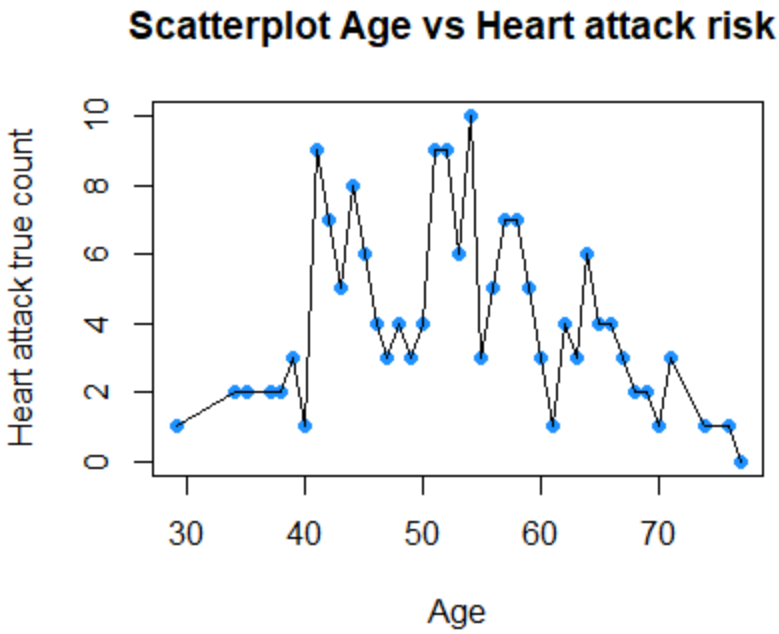
Como podemos ver en la gráfica hay un mayor número de pacientes que tienen azúcar en sangre en ayunas igual o por debajo de 120 mg/dl, y dentro de este conjunto vemos que sucede en mayor proporción en pacientes de género *sex* = 1.

Como podemos ver en ambas gráficas hay mayor proporción de pacientes de género *sex* = 1.

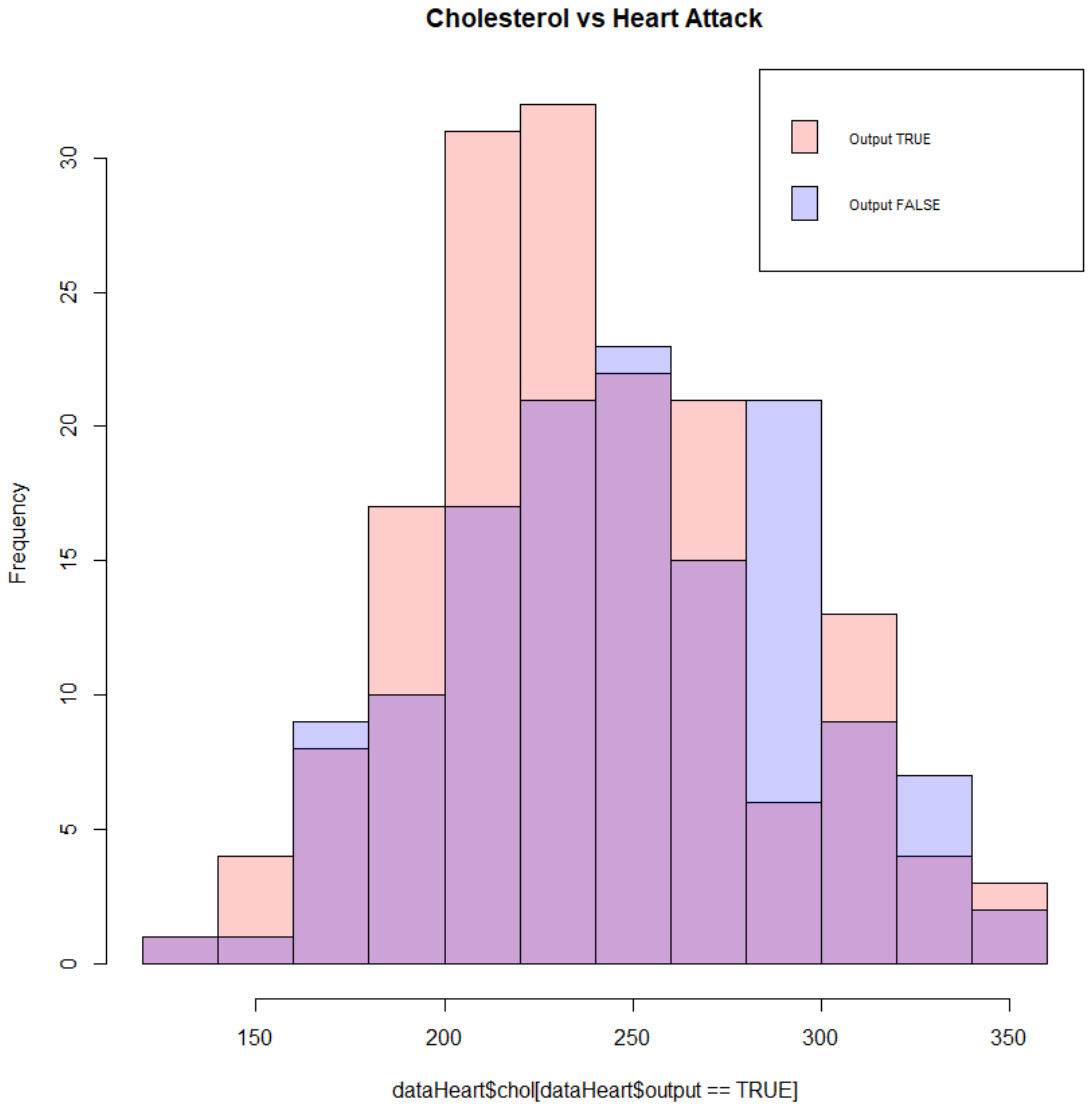
```
hist(dataHeart$thalach [dataHeart$sex==TRUE], col=rgb(1,0,0,0.2))
hist(dataHeart$thalach [dataHeart$sex==FALSE], col=rgb(0,0,1,0.2), add=TRUE)
legend('topright', c('Sex TRUE', 'Sex FALSE'),
       fill=c(rgb(1,0,0,0.2),rgb(0,0,1,0.2)), xpd=TRUE, cex=0.7,)
```



Como se puede ver en el histograma en ambos sexos siguen una distribución muy parecida, en ambos casos los picos más altos se concentran entre 140 y 175 la frecuencia cardíaca máxima alcanzada.



Se puede observar que los ataques del corazón son menos comunes entre la gente menor de 40 años, sin embargo y a pesar de lo que presenta la gráfica no podemos decir que la probabilidad de la gente mayor de 70 sea menor debido al pequeño número de muestras de esta población.



Cholesterol vs output:
Como podemos observar en esta gráfica, parece haber un pico en el número de infartos cuando el colesterol supera los 200. Por ello vamos a separar dos grupos en función de si el colesterol es o no alto.

```
dataHeartCholH <- dataHeart[dataHeart$chol >= 200,]
dataHeartCholL <- dataHeart[dataHeart$chol < 200,]
```

Por ello vamos a proceder a realizar un contraste de hipótesis para observar si la media de ritmo cardíaco máximo es mayor entre personas con el colesterol alto (mayor de 200).

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

```
library(nortest)
alpha = 0.05
col.names = colnames(dataHeart)
for (i in 1:ncol(dataHeart)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(dataHeart[,i]) | is.numeric(dataHeart[,i])) {
    p_val = ad.test(dataHeart[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      # Format output
      if (i < ncol(dataHeart) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

Para comprobar la normalidad de las variables numéricas hemos usado el método de Anderson-Darling.

```
Variables que no siguen una distribución normal:
age, trtbps, thalachh, oldpeak,
```

Las variables que se muestran son las que siguen una distribución de normalidad.

Hacemos una comprobación de la homogeneidad de la varianza de la muestra de posterior estudio:

```
dataHeart$Hcol <- cut(dataHeart$chol, breaks = c(0,200,1000), labels =
c(0,1))
fligner.test(thalachh ~ Hcol, data = dataHeart)
```

```
      Fligner-Killeen test of homogeneity of variances

data:  thalachh by Hcol
Fligner-Killeen:med chi-squared = 3.0328, df = 1, p-value = 0.0816
```

Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

¿La frecuencia es mayor si el colesterol es alto?

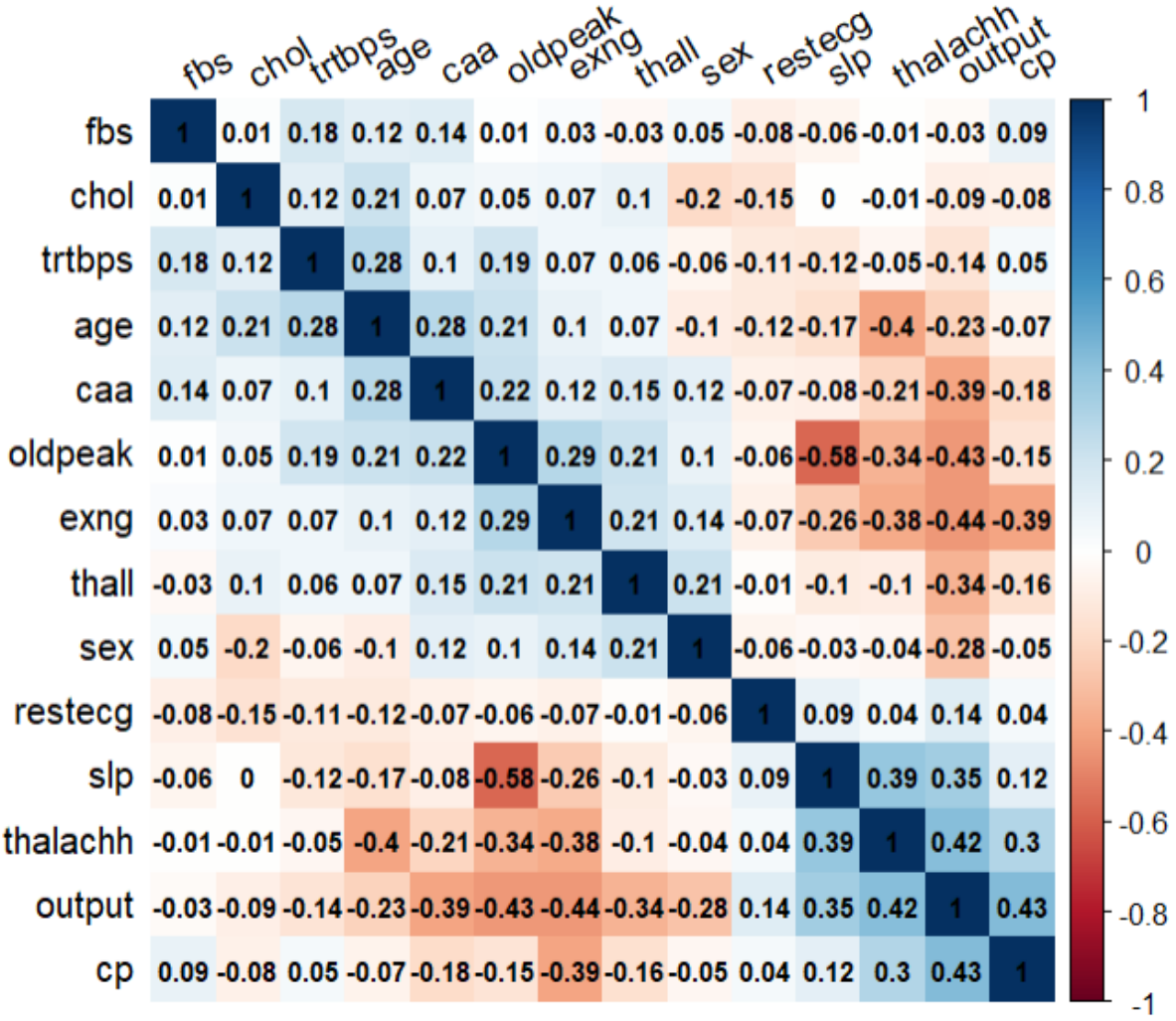
Hipótesis Nula:
La media de frecuencia cardiaca máxima es igual en las personas con el colesterol alto (Mayor de 200).

Hipotesis alternativa:
La media de frecuencia cardiaca máxima es menor en las personas con el colesterol alto (Mayor de 200).

```
t.test(dataHeartCholL$thalachh, dataHeartCholH$thalachh, alternative =
"less")
```

```
      Welch Two Sample t-test

data:  dataHeartCholL$thalachh and dataHeartCholH$thalachh
t = -0.33218, df = 65.128, p-value = 0.3704
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 5.140487
sample estimates:
mean of x mean of y
 148.5800  149.8577
```

En este caso observamos un p-value de 0.37 por lo que no podemos rechazar la hipótesis nula al tener p value mucho mayor al valor de significación fijado. Por tanto no se observan diferencias significativas entre las medias (148.5800 149.8577).



Como se puede observar en la matriz de correlación existe una cierta correlación positiva significativa entre la variable *output* y *chest pain (cp),* y entre las variables *thalachh* y *slp*. Por otro lado, existe una correlación negativa entre las variables *slp* y *oldpeak*.

Ya que nuestra variable independiente, *output*, es dicotómica, el modelo que se aplicará será el modelo de regresión logística.

```
ModlgF <- glm(output ~ age+sex+cp+trtbps+chol+fbs+restecg+thalachh+exng+oldpeak+slp+
caa+thall+output, data = dataHeart, family = "binomial")
summary(ModlgF)
```

```
> summary(ModlgF)

Call:
glm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
    restecg + thalachh + exng + oldpeak + slp + caa + thall +
    output, family = "binomial", data = dataHeart)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5849  -0.3872   0.1551   0.5863   2.6249

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.450472   2.571479   1.342 0.179653
age         -0.004908   0.023175  -0.212 0.832266
sex         -1.758181   0.468774  -3.751 0.000176 ***
cp           0.859851   0.185397   4.638 3.52e-06 ***
trtbps      -0.019477   0.010339  -1.884 0.059582 .
chol        -0.004630   0.003782  -1.224 0.220873
fbs          0.034888   0.529465   0.066 0.947464
restecg      0.466282   0.348269   1.339 0.180618
thalachh     0.023211   0.010460   2.219 0.026485 *
exng        -0.979981   0.409784  -2.391 0.016782 *
oldpeak     -0.540274   0.213849  -2.526 0.011523 *
slp          0.579288   0.349807   1.656 0.097717 .
caa         -0.773349   0.190885  -4.051 5.09e-05 ***
thall       -0.900432   0.290098  -3.104 0.001910 **
---
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: 211.44  on 289  degrees of freedom
AIC: 239.44

Number of Fisher Scoring iterations: 6
```

Viendo el resumen del modelo vemos que las variables *age*, *trtbps*, *chol*, *fbs*, *restecg* y *slp* no son significativas ya que el valor de p-value es superior a 0.05.

5. Representación de los resultados a partir de tablas y gráficas. Este apartado se ha resuelto a lo largo de la práctica.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

No hay valores nulos o vacíos en los datos.

Se presentan ciertos valores atípicos en las variables cuantitativas, sin embargo estos parecen tener sentido para cada una de las variables estudiadas y solo se han descartado los outliers de colesterol.

No parece haber una correlación directa entre las variables continuas según el estudio del correlation plot.

Tenemos pocos datos para saber si es más probable tener un ataque al corazón según la edad, sin embargo la tendencia parece ser que es menos probable en menores de 40 y mayores de 70 (a pesar de que esto resulte contra intuitivo).

Si nos fijamos en el resumen que nos proporciona el modelo de regresión logística vemos que las características *cp*, *fbs*, *restecg*, *thalachh* y *slp* son factores de riesgo, es decir, que la presencia de estas variables dentro del modelo favorecen al suceso de padecer un ataque al corazón.

Por otro lado las variables *age*, *sex*, *trtbps*, *chol*, *exng*, *oldpeak*, *caa* y *thall* son factores de protección, esto quiere decir, que en presencia de estas características el suceso de padecer un ataque al corazón es menos probable.

Por tanto, un paciente con una frecuencia cardiaca máxima (*thalachh*) mayor tendrá mayores probabilidades de sufrir un ataque al corazón debido a su correlación y significancia con output.

Asimismo, los resultados de los electrocardiogramas (*restecg*) parecen resultar también útiles a la hora de predecir un ataque al corazón debido a la significancia de los mismos.

Del mismo modo, en los datos se observa un pico en el número de casos de ataques al corazón cuando observamos un colesterol por encima de 200 de forma visual.

Sin embargo, no se han encontrado diferencias significativas entre las medias de frecuencia cardiaca máxima entre las personas con colesterol por encima de 200 respecto a las que lo tienen menor por contraste de hipótesis. *Por tanto, no se observa una diferencia significativa entre los ritmos cardiacos máximos según el nivel de colesterol.*

Para obtener estos resultados se han aplicado distintas técnicas como pueden ser el análisis visual, el contraste de hipótesis, la regresión logística y correlación entre variables.

7. Código. Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Se ha trabajado sobre el código en R comentado en este documento, en este caso se puede acceder a este en el siguiente enlace:
https://github.com/DanielRomeroResano/PRAC2-Tipolog-a/tree/main/code

8. Vídeo.

Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace al Google Drive de la UOC (https://drive.google.com/…), junto con enlace al repositorio Git entregado.

https://github.com/DanielRomeroResano/PRAC2-Tipolog-a

https://drive.google.com/file/d/1ahdWO9ESS-a698AQSyagCnFaT809paEV/view?usp=share_link

Contribuciones	Firma
Investigación previa	Fabrizio, Daniel RR
Redacción de las respuestas	Fabrizio, Daniel RR
Desarrollo del código	Fabrizio, Daniel RR
Participación en el vídeo	Fabrizio, Daniel RR