Práctica 2 – ¿Cómo realizar la limpieza y el análisis de los datos?

Heart Attack Analysis & Prediction Dataset

Aysha Ait Ouaddi El Mamouny

Contents

Se realiza la importación del conjunto de datos ***Heart Attack Analysis & Prediction Dataset*** [1]:

```
df=read.csv("D:\\43591894v\\Downloads\\heart.csv", sep=',')
```

A continuación, se realiza una primera inspección de los datos mediante la examinación de la tipologia de cada variable y la realización de un análisis descriptivo sencillo.

```
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
```

1. Descripción del dataset

El conjunto de datos que se ha escogido para la realización de la práctica es el *Heart Attack Analysis & Prediction Dataset* de Kaggle, el cual recoge datos relativos a casos en que se han producido ataques cardiovasculares. Por lo tanto, consta de las siguientes variables [2]:

- **Age**: Edad del paciente.
- **Sex**: Sexo del paciente. (1 = Hombre; 0 = Mujer)
- **Exng**: Angina inducida por ejercicio. (1 = Sí; 0 = No)
- **Caa**: Número de vasos principales. (0-3)
- **Cp**: Tipo de dolor torácico. (0 = Angina típica; 1 = Angina atípica; 2 = Dolor no anginoso; 3 = Asintomático)
- **Trtbps**: Tensión 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. (0 = Normal; 1 = Con anomalía de la onda ST-T (inversión de la onda T y/o elevación o depresión del ST > 0,05 mV); 2 = Hipertrofia ventricular izquierda probable o definida según los criterios de Estes)
- **Thalachh**: Frecuencia cardiaca máxima alcanzada.
- **Output**: Probabilidad de padecer un infarto. (0 = Menos posibilidades de infarto; 1 = Más posibilidades de infarto)
- **Oldpeak**: Depresión del ST inducida por el ejercicio en relación con el reposo.
- **Spl**: Pendiente del pico del segmento ST de ejercicio. (1 = Asciendiente; 2 = Equilibrado; 3 = Descendiente)
- **Thall**: Grado de talasemia: (1 = Defecto fijo; 2 = Normal; 3 = Defecto reversible)

NOTA: ST hace referencia a las posiciones en el gráfico del electrocardiograma.

Se trata de un conjunto de datos de sumo interés ya que abre la posibilidad a responder preguntas relativas a las circunstancias en que se producen ataques cardíacos. Algunos ejemplos de las múltiples preguntas que se pueden realizar en este contexto son:

- ¿Existe mayor predisposición a padecer un ataque cardiovasuclar en función del sexo?
- ¿Existen variables correlacionadas?
- ¿Es posible determinar que factores fomentan la probabilidad de padecer un ataque cardíaco
- ¿Se puede pronosticar la probabilidad de padecer dicha patología?

Finalmente, cabe mencionar que se trata de un conjunto de datos que procede de datos médicos reales, lo cual indica que las conclusiones extraídas serán aproximadas a la realidad.

2. Integración y selección

No se realizarán tareas de integración ya que el conjunto de datos original dispone de la variedad de datos necesarios para el objetivo del estudio, por lo tanto, no se necesita combinar los datos con otros procedentes de otras fuentes. De la misma manera, tampoco se realizarán tareas de subselección o filtrado ya que para el objeto de este estudio, se desea trabajar con el conjunto de datos original completo ya que es muy variado.

3. Limpieza de los datos

3.1. ¿Los datos contienen ceros o elementos vacíos?

A continuación, es preciso determinar si existen valores nulos en las variables del dataset. Tras examinar los datos se obtiene el siguiente resultado, lo cual indica que no existen nulos en los datos y consecuentemente no se precisa realizar tareas de preprocesado por lo que se refiere a este aspecto:

```
# Se obtiene el total de nulos por variable y se imprime el resultado en una tabla.
library(knitr)
res=colSums(is.na(df))
tabla <- data.frame(atributo = names(res),nulos = res);kable(tabla, row.names = F)
```

atributo	nulos
age	0
sex	0

atributo	nulos
cp	0
trtbps	0
chol	0
fbs	0
restecg	0
thalachh	0
exng	0
oldpeak	0
slp	0
caa	0
thall	0
output	0

Por otro lado, pese a que todas las variables son originalmente numéricas, existen variables que tienen una connotación categórica. Estas variables son: *sex*, *cp*, *fbs*, *restecg*, *slp*, *caa*, *thall*, *output* y *exng*. Consecuentemente, se convertirán a tipo factor:

```
# Se realiza a conversión de las variables indicadas en la lista a tipo factor.
a_factor <- c(2,3,6,7,9,11,12,13,14)
df[a_factor] <- purrr::map_df(df[a_factor], as.factor)
```

3.2. Identifica y gestiona los valores extremos

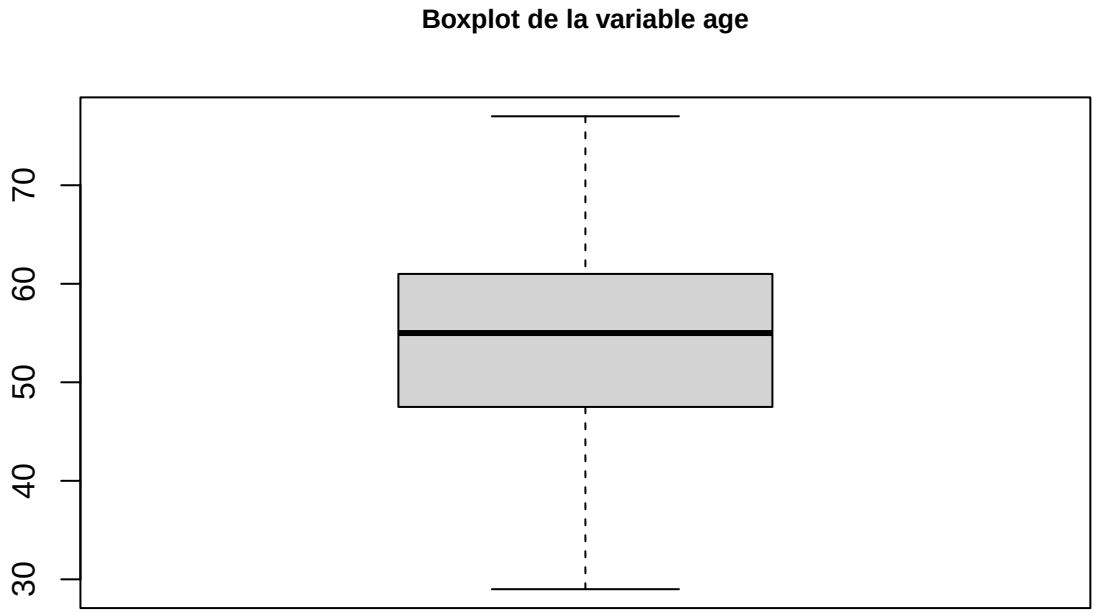
Por otro lado, es posible determinar los valores extremos de las variables numéricas mediante el uso de boxplots o el uso de la función propia de R *boxplots.stats()*. Por lo tanto, primeramente se aplicará la función anteriormente mencionada a cada una de las variables numéricas para determinar explícitamente los valores extremos.

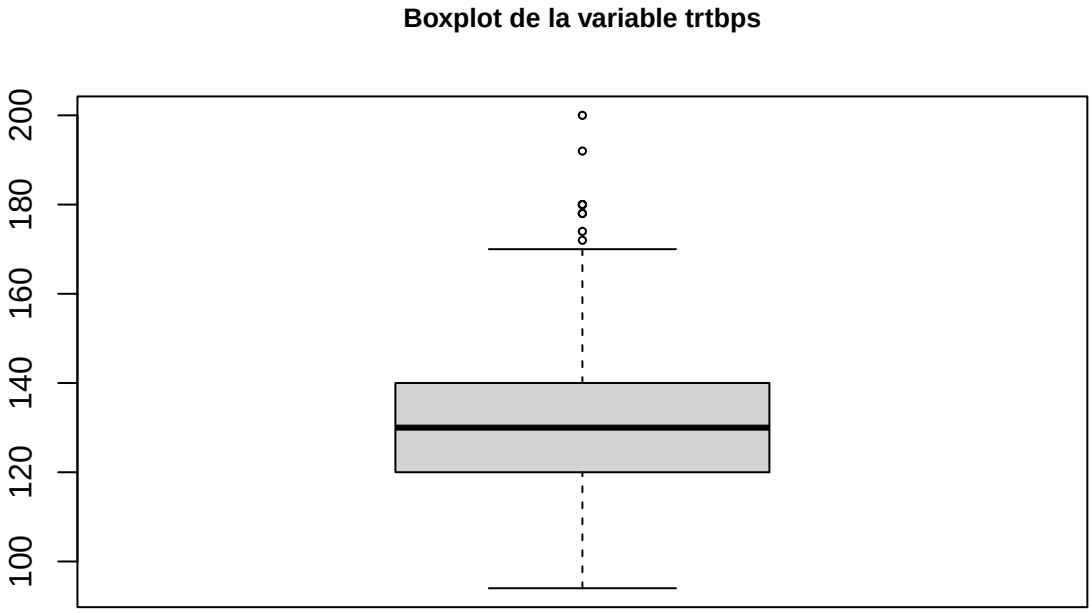
```
# Mediante un bucle se imprimen en una tabla los outliers de las variables numéricas.
resultados <- data.frame(atributo = character(), outliers = character())

for (i in names(df)) {
  if (is.numeric(df[,i]) || is.integer(df[,i])) {
    outliers <- boxplot.stats(df[,i])$out
    resultados <- rbind(resultados, data.frame(atributo = i, outliers = paste(outliers, collapse = ", "
  }
}

kable(resultados, row.names = FALSE)
```

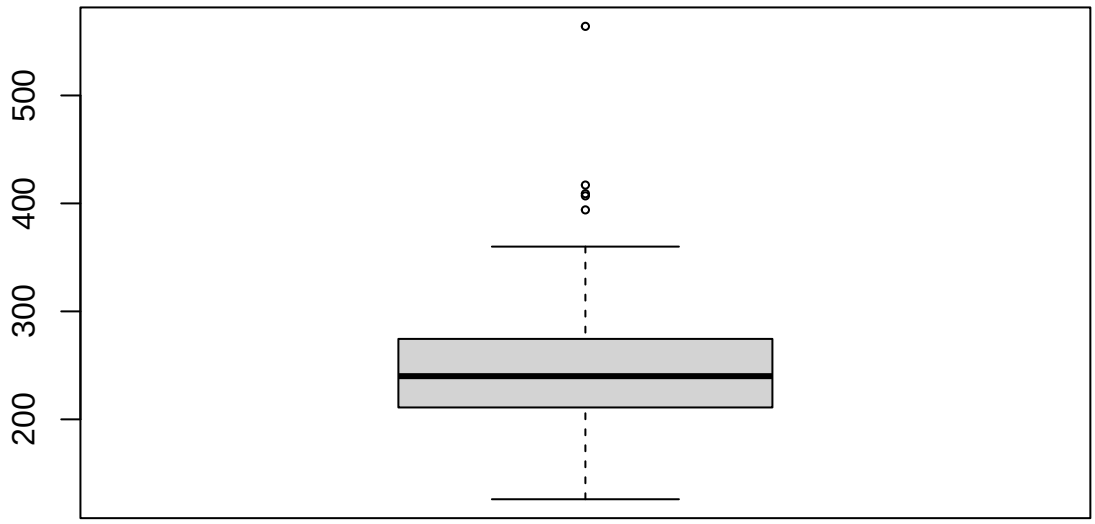
atributo	outliers
age	
trtbps	172, 178, 180, 180, 200, 174, 192, 178, 180
chol	417, 564, 394, 407, 409
thalachh	71
oldpeak	4.2, 6.2, 5.6, 4.2, 4.4

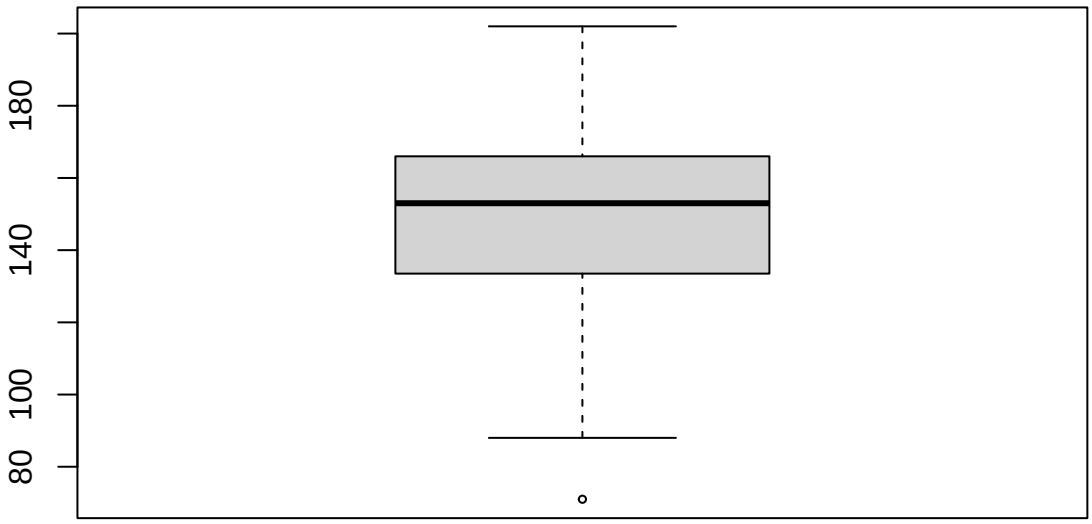
Adicionalmente, es recomendable estudiar visualmente el comportamiento de los valores extremos utilizando boxplots para cada una de las variables numéricas.

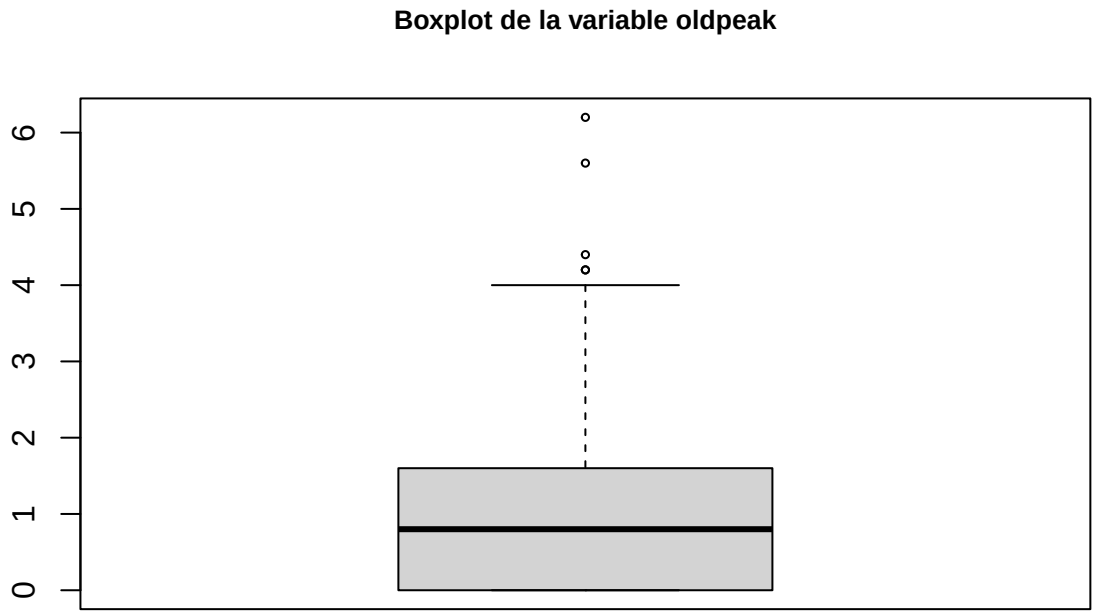
```r
# Mediante un bucle se visualizan los boxplots de cada una de las variables numéricas.
for (i in names(df)) {
  if (is.numeric(df[,i]) || is.integer(df[,i])) {
    boxplot(df[,i], main =paste("Boxplot de la variable",i), cex.main = 0.8, cex=0.5)
  }
}
```




Boxplot de la variable trtbps
100
120
140
160
180
200


Boxplot de la variable chol
200
300
400
500


Boxplot de la variable thalachh
80
100
140
180



Realizado el procedimiento anterior, es necesario verificar si los valores atípicos detectados son anormales. A continuación, se comentan los valores atípicos de cada una de las variables en función del contexto:

- *Age*: No se detectan valores atípicos.
- *Trtbps*: Se considera que los valores detectados no son anormales dado que existen casos puntuales de hipertensión grave que pueden provocar valores superiores a 170 mm Hg. [3]
- *Chol*: Se considera que los valores detectados no son anormales dado que existen casos relacionados con hipertrigliceridemia e hipercolesterolemia que pueden provocar valores como los observados. [4]
- *Thalachh*: Se considera que el valor detectado no es anormal dado que se trata de un valor que puede llegar a ser observado en la población. [5]
- *Oldpeak*: Se considera que los valores detectados no son anormales dado que existen casos relacionados a episodios de infartos que pueden provocar valores como los observados. [6]

Por lo tanto, no se realizarán tareas de procesado por lo que se refiere a este aspecto dado que se trata de outliers que no deberían ser eliminados del conjunto de datos.

4. Análisis de los datos

4.1. Selección de los grupo de datos

En función de las pruebas estadísticas que se aplicarán en el apartado *4.3. Aplicación de pruebas estadísticas* se crearán grupos de datos comparables.

En primer lugar, dado que se pretende analizar la correlación entre variables puramente numéricas mediante una matriz, se realizará una selección que incluya únicamente dichas variables:

```
# Se seleccionan únicamente las variables numéricas del dataset.
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
var_numericas <- select_if(df, is.numeric)
```

Adicionalmente, se estudiará la correlación entre las variable independientes tipo factor y la variable objetivo, se creará una selección que incluya únicamente estas primeras.

```
# Se crea un conjunto de datos con todas las variables independientes.
variables <- names(df)
variables <- variables[variables != "output"]
```

En segundo lugar, se analizará la edad en función del sexo por lo cual se compararán los grupos de datos siguientes (NOTA: En el punto 4.3 no se hará uso de las variables creadas):

```
# Se crean dos conjuntos de datos referentes a la edad por sexos.
edades_mujeres <- subset(df, sex == 0, select = age)
edades_hombres <- subset(df, sex == 1, select = age)
```

En tercer lugar, se analizará la angina inducida por ejercicio en función de tipo de dolor torácico (variables categóricas):

```
# Se crea una tabla de contingencia con las variables *exng* y *cp*.
cp_exng=table(df$exng,df$cp)
```

En cuarto lugar, se creará un modelo de regresión logística para predecir la posibilidad de que un paciente tenga un infarto. Para ello, se creará un conjunto de entrenamiento con el 80% de los datos y un conjunto de prueba con el 20% restante.

```
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
indices <- createDataPartition(df$output, p = 0.8, list = FALSE)
# Se crea un conjunto de entrenamiento y un conjunto de test.
entrenamiento <- df[indices,]
prueba <- df[-indices,]
```

Finalmente, será objeto de estudio la tensión arterial en reposo en función del tipo de dolor torácico, por lo que se compararán los siguientes grupos (NOTA: En el punto 4.3 no se hará uso de las variables creadas):

```
# Se crean sub-conjuntos de datos de la variable *trtbps* en función de *cp*.
tensión_angina_tipica <- subset(df, cp ==0 , select = trtbps)
tension_angina_atípica <- subset(df, cp == 1, select = trtbps)
tension_no_anginoso <- subset(df, cp == 2, select = trtbps)
tension_asintomatico <- subset(df, cp == 3, select = trtbps)
```

4.2. Comprobación de la normalidad y homogeneidad de la varianza

Existen múltiples test caracterizados por diferentes enfoques matemáticos que permiten realizar la comprobación de la normalidad de las variables. En este caso en particular, se hara uso del *Shapiro-Wilk test*, que asume como hipótesis nula que la población está distribuida normalmente. [7]

Tras realizar el procedimiento, el test indica que ninguna variable es normal, ya que el p-valor es inferior al coeficiente de significación 0.05, por lo que se puede rechazar la hipotesis nula.

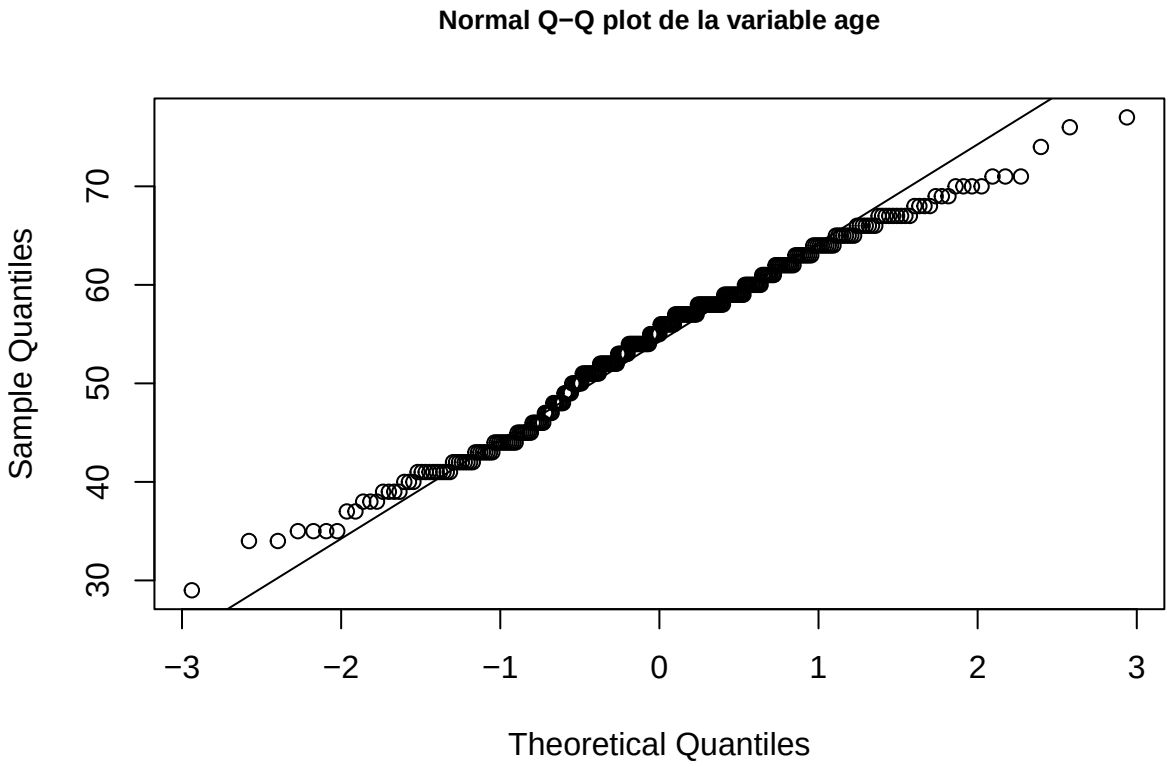
```
numeric_columns <- sapply(df, is.numeric)

# Se seleccionan solo las columnas numéricas del dataframe.
data_numeric <- df[, numeric_columns]

# Se aplica la prueba de Shapiro-Wilk a cada columna.
p_values <- apply(data_numeric, 2, shapiro.test)

# Se obtienen los nombres de los atributos y los p-valores en dos vectores separados
atributos <- names(p_values)
p_values <- sapply(p_values, function(x) x$p.value)

# Se unen los dos vectores en una tabla con la función cbind
results_table <- cbind(atributos, p_values)

# Se imprime la tabla con kable
kable(results_table, col.names = c("Atributo", "p-valor"), align = c("l", "r"), row.names = F)
```

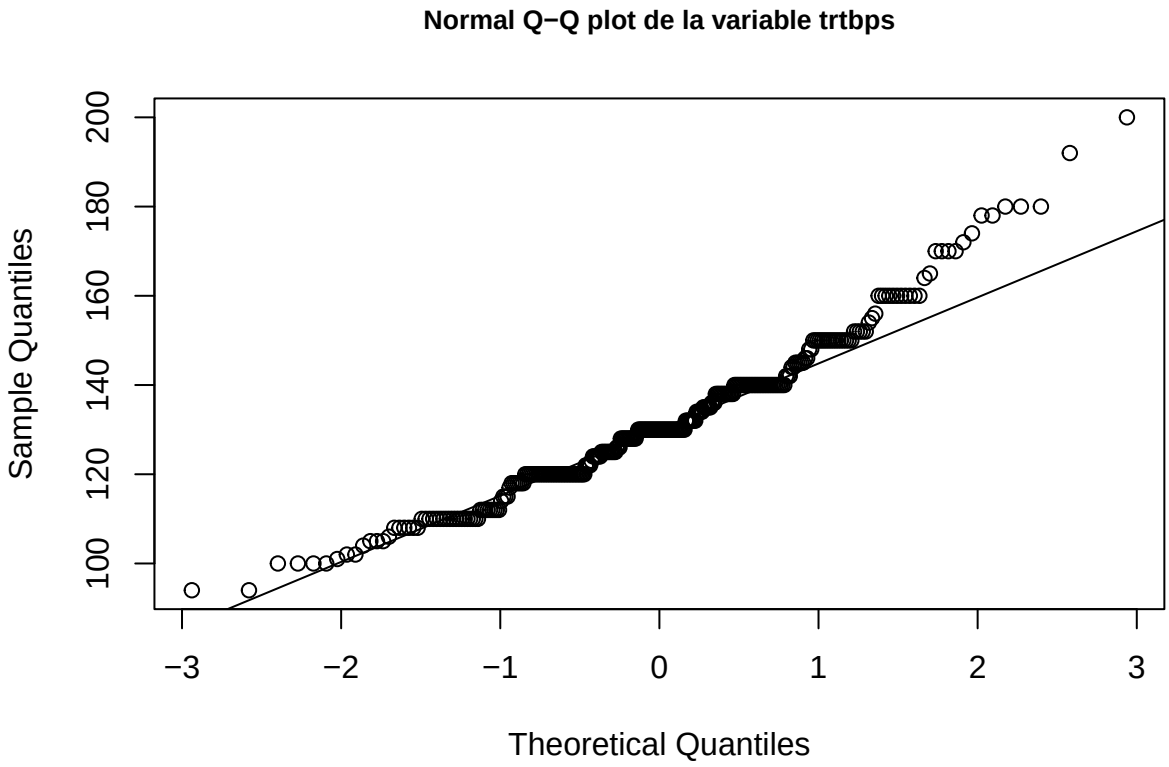
Atributo	p-valor
age	0.00579835873933943
trtbps	1.45809705294778e-06
chol	5.36484787285249e-09
thalachh	6.62081936641719e-05
oldpeak	8.18337828923442e-17

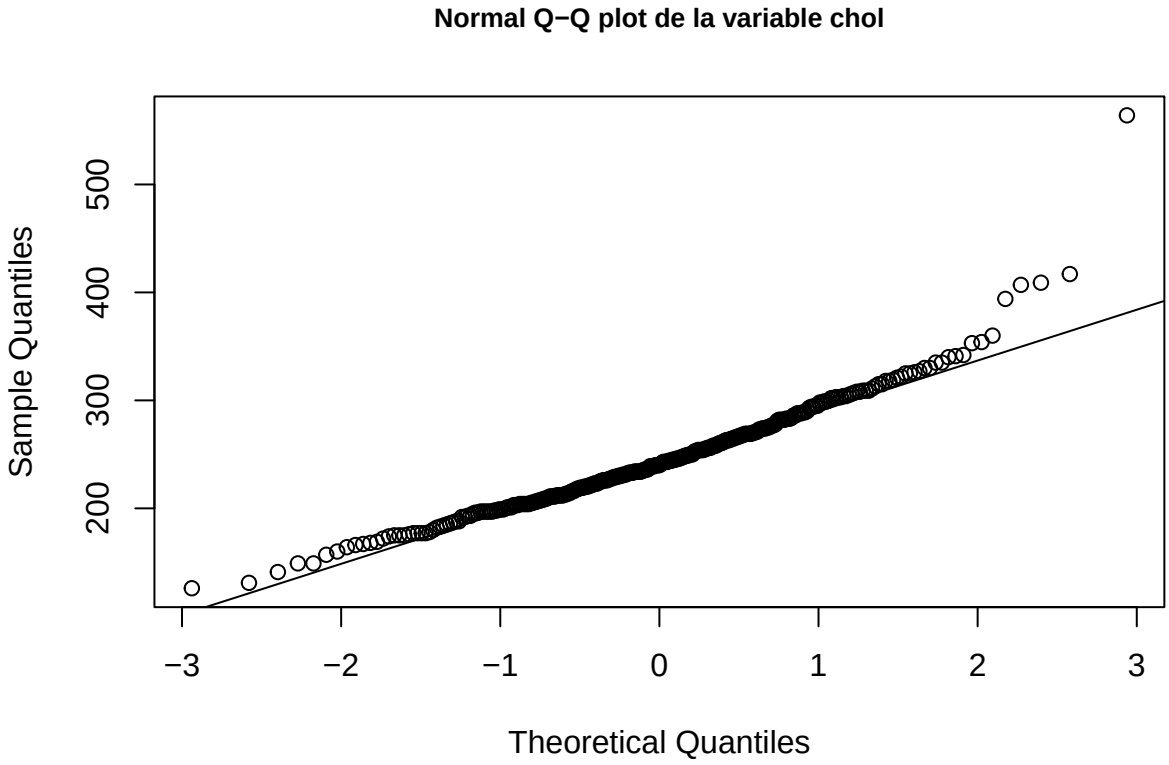
Adicionalmente, es posible examinar este hecho de manera visual mediante el uso de gráficos de Q-Q (quantile-quantile). Realizadas estas respresentaciones. De esta manera es posible observar que las variables pueden tender a una distribución normal.

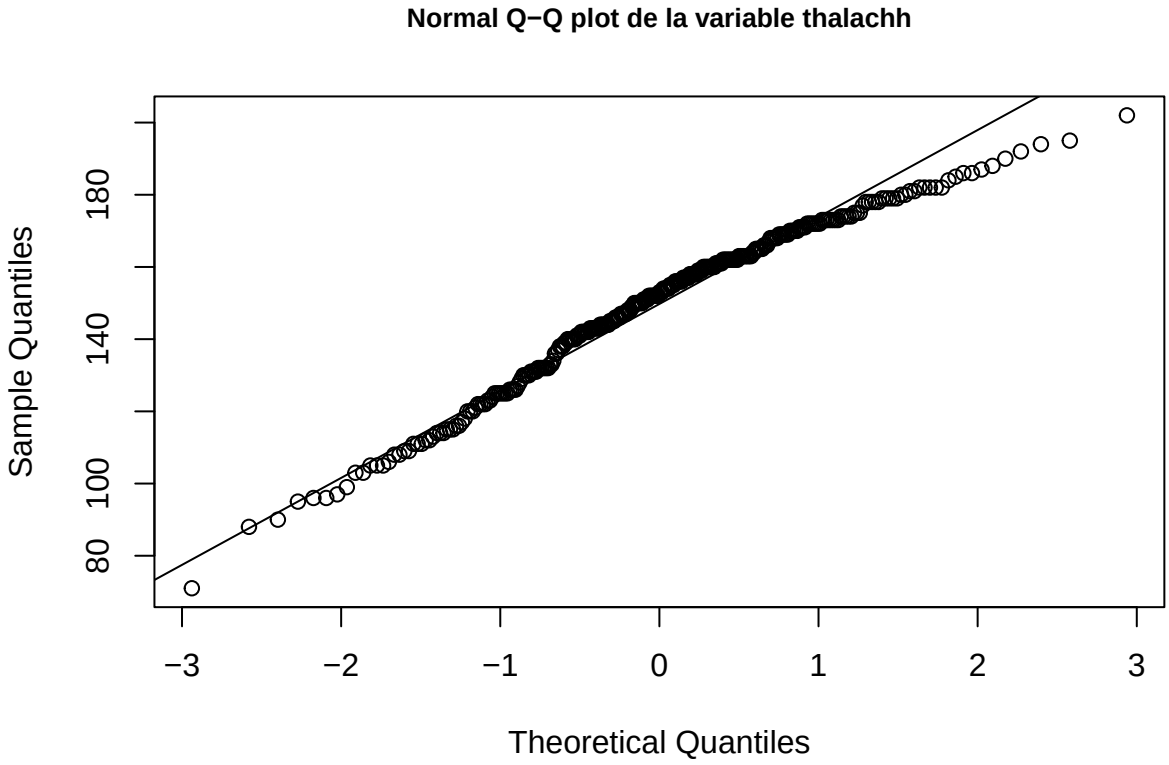
```
# Se crea un bucle para imprimir los gráficos de Q-Q de cada una de las variables numéricas.
numeric_cols <- sapply(df, is.numeric)
int_cols <- sapply(df, is.integer)
selected_cols <- names(numeric_cols)[numeric_cols | int_cols]
```

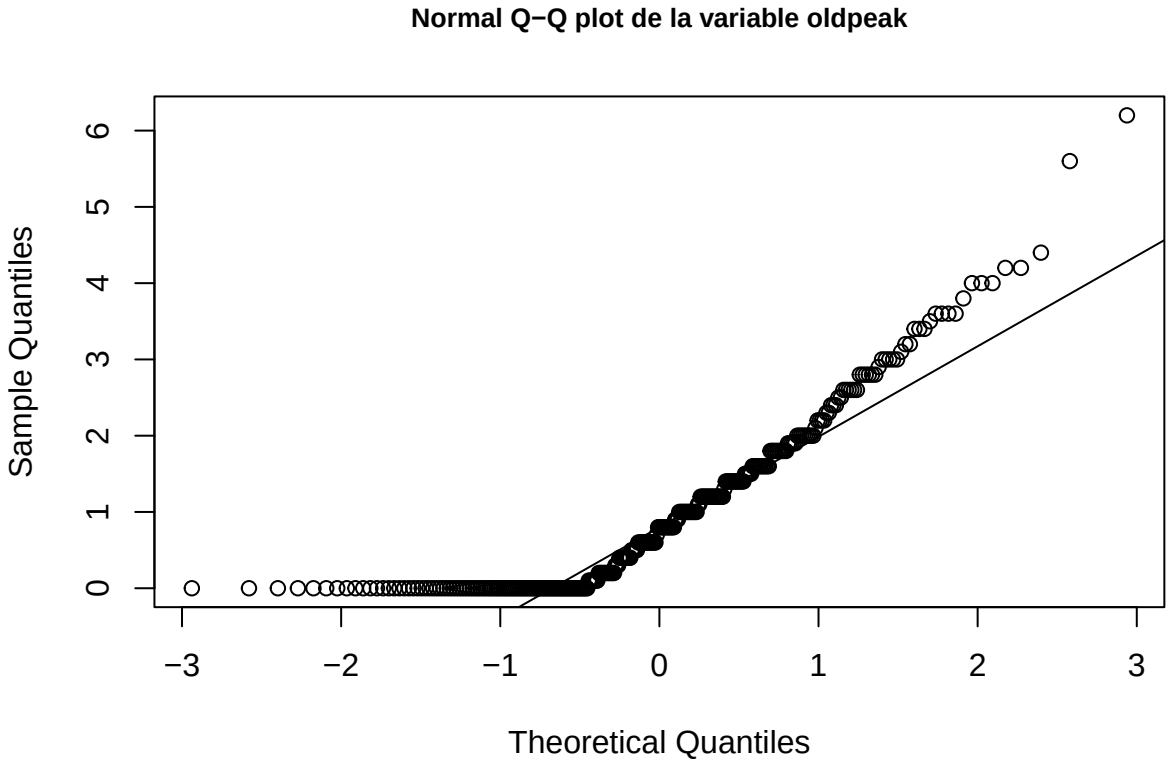
```r
plots <- lapply(selected_cols, function(x) {
  qqnorm(df[[x]], main="")
  qqline(df[[x]])
  title(main=paste("Normal Q-Q plot de la variable",x), cex.main=0.8)
})
```




Normal Q−Q plot de la variable trtbps
Sample Quantiles
100
120
140
160
180
200
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q plot de la variable chol
Sample Quantiles
200
300
400
500
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q plot de la variable thalachh
Sample Quantiles
80
100
140
180
−3
−2
−1
0
1
2
3
Theoretical Quantiles




```
invisible(plots)
```

Sin embargo, según el teorema del limite central al disponer de más de 30 elementos en las observaciones es posible aproximar las variables a una distribución normal de media 0 y desviación estandard 1. [7]

Por otro lado, dado que se plantea la realización de pruebas de comparación entre grupos *(véase los puntos 4.3.3 y 4.3.6)*, es preciso estudiar la homogeneidad de las varianzas. Con este objeto, y dado que se considera que la distribución de la variable *age* y *trtbps* son normales, se realiza el *Levene Test*. En esta prueba, 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. Por lo tanto, una vez realizado el procedimiento, se obtiene que los p-valor son superiores al nivel de significancia 0.05, por lo cual se acepta la hipótesis nula de que la varianza entre los grupos es igual.

```
# Se aplica el test de Levene a *age* en función de *sex*.
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

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

```r
levene_test <- leveneTest(age ~ sex, data=df); levene_test
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1   0.363 0.5473
##       301
```

```r
# Se aplica el test de Levene a *trtbps* en función de *cp*.
levene_test <- leveneTest(trtbps ~ cp, data=df); levene_test
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  1.4061 0.2411
##       299
```

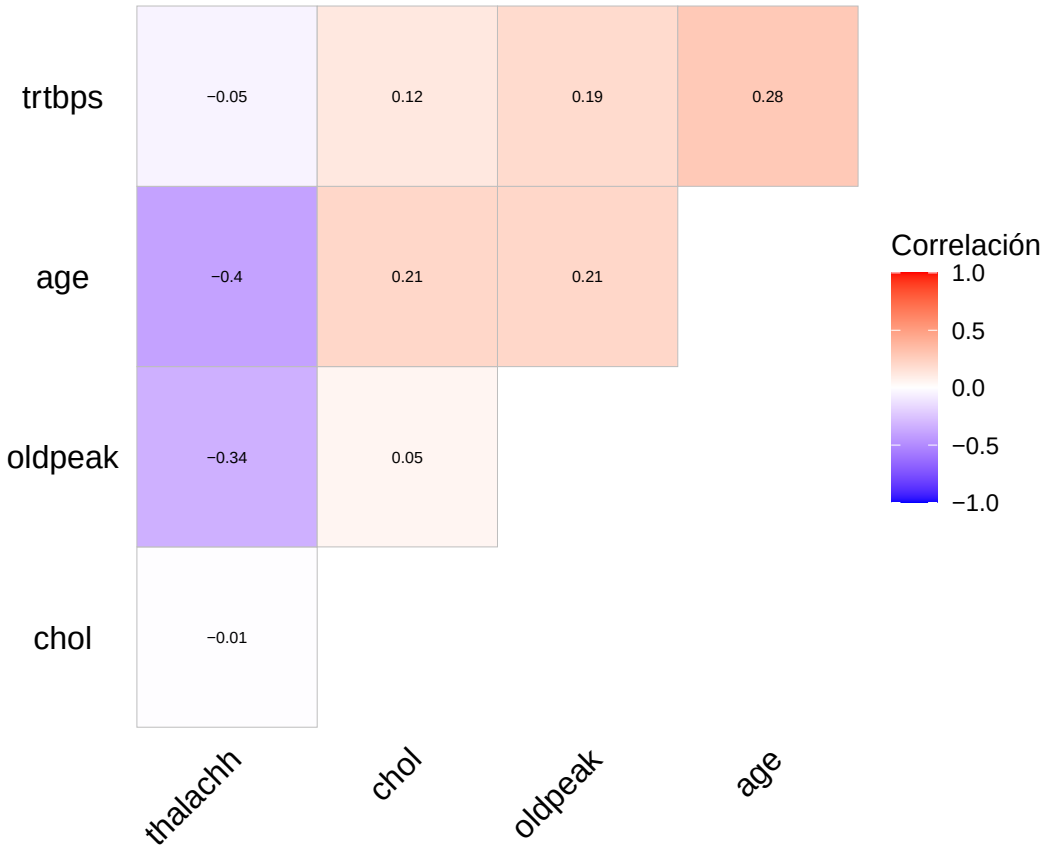
4.3. Aplicación de pruebas estadísticas

4.3.1. Estudio de correlación lineal entre variables numéricas.

Resulta de interés averiguar si existe relación lineal entre las variables numéricas disponibles en el conjunto de datos. Por lo tanto, se determinará este hecho haciendo uso de una matriz de correlación de pearson, ya que se ha verificado el criterio de normalidad.

Observados los resultados de la matriz, se puede determinar que no existe una correlación lineal especialmente fuerte entre ellas. Pese a esto, cabe mencionar la ligera correlación negativa existente entre la frecuencia cardiaca máxima alcanzada y la edad, así como con la depresión del ST inducida por el ejercicio. También se observan muy ligeras correlaciones entre la edad y el colesterol, la edad y Depresión del ST inducida por el ejercicio o la edad y la tensión arterial.

```r
# Se realiza una matriz de correlación con las variables numéricas usando la función *ggcorrplot*.
library(ggcorrplot)
ggcorrplot(cor(var_numericas, method="pearson"), method = "square", hc.order = TRUE, lab=TRUE, lab_col =
```



4.3.2. Estudio de diferencias significativas entre las variables categóricas y la variable objetivo

Por otro lado, conviene realizar un análisis de correlación que permita incluir variables categóricas, con el fin de estimar qué variables tienen relación con *output* (la probabilidad de padecer un ataque o no) es determinar el p-valor que vincula cada una de ellas con *output*. Para ello, dado que se trata de variables tipo factor es recomendable para este caso hacer uso del *Test Chi-Cuadrado*. [7]

```
# Se inicializa una tabla vacía.
tabla <- data.frame(matrix(ncol = 2, nrow = 0))

# Se añaden los nombres de columna a la tabla
colnames(tabla) <- c("Variable", "p-valor")

variables <- names(df)
variables <- variables[variables != "output"]
for (variable in variables) {
  if(is.factor(df[, variable])) {
  # Se calcula la prueba de chi.cuadrado de independencia entre cada variable categórica y la variable
  resultado <- chisq.test(df[, variable], df$output)

  # Se añade el resultado a la tabla.
  tabla <- rbind(tabla, c(variable, resultado$p.value))
}}
```

```
## Warning in chisq.test(df[, variable], df$output): Chi-squared approximation may
```

```
## be incorrect

## Warning in chisq.test(df[, variable], df$output): Chi-squared approximation may
## be incorrect

## Warning in chisq.test(df[, variable], df$output): Chi-squared approximation may
## be incorrect
```

```
colnames(tabla) <- c("Atributo", "p-valor")


# Se muestra la tabla usando la función kable.
kable(tabla)
```

Atributo	p-valor
sex	1.87677762169415e-06
cp	1.33430433730501e-17
fbs	0.744428111414958
restecg	0.00666059877349803
exng	7.45440933123567e-14
slp	4.83068193427684e-11
caa	2.71247021195932e-15
thall	2.23335072101293e-18

Observados los resultados, el test entre la variable objetivo y *fbs* arroja un p-valor de 0.74 lo cual induce a aceptar la hipótesis nula y se puede concluir que no hay evidencia suficiente para rechazar la hipótesis nula. Esto significa que no podemos decir con certeza que existe una diferencia significativa entre *output* y *fbs*. Por contra, el p-valor del resto de variables hace rechazar la hipótesis nula y concluir que existen diferencias significativas entre ellas y la variable objetivo.

4.3.3. Prueba t de Student

En este apartado se desea estudiar si existen diferencias significativas por lo que se refiere a la edad en función del sexo. Dado que previamente se ha comprobado la normalidad de la variable *age* y homoscedasticidad de los grupo conformados por hombres y mujeres, se hará uso de la prueba *t de Student*. En esta prueba, la hipótesis nula asume que las medias de los grupos de datos son las mismas.

Observado el resultado que proporciona el p-valor, el cual supera el nivel de significación 0.05, se concluye que no existen diferencias estadísticamente significativas. También, cabe mencionar que mediante el test se pueden observar las medias de ambas poblaciones, las cuales no son muy diferentes, hecho que concuerda con la hipótesis contrastada.

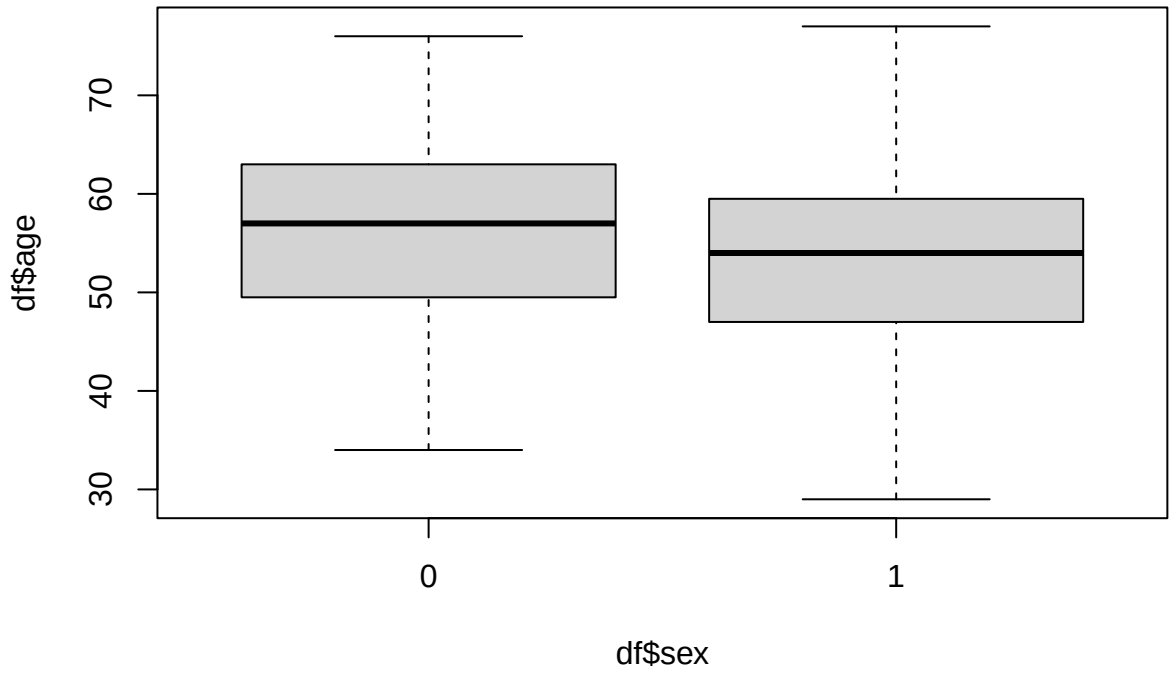
```
# Se realiza un t-test teniendo en cuenta las variables *age* y *sex*.
t.test(age ~ sex, data=df)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by sex
## t = 1.6805, df = 175.92, p-value = 0.09464
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.3346005  4.1718589
## sample estimates:
## mean in group 0 mean in group 1
##        55.67708        53.75845
```

Asimismo, esto también se puede observar visualmente mediante un boxplot que contenga las edades por sexo.

```
# Se realiza un boxplot teniendo en cuenta las variables *age* y *sex*.
boxplot(df$age  ~ df$sex)
```



4.3.4. Prueba Chi-cuadrado de Pearson

En este apartado se desea analizar si existen diferencias significativas entre la angina inducida por ejercicio y el tipo de dolor torácico, ambas variables categóricas [7]. En estos casos, es posible aplicar la prueba Chi-cuadrado.

Observado el resultado del test, se puede concluir que existe una diferencia estadísticamente significativa entre las variables categóricas comparadas, lo cual significa que es poco probable que la diferencia observada entre los grupos o variables se deba al azar o a la variabilidad natural de la muestra.

```
# Se realiza el test Chi-Cuadrado con la tabla de contingencia.
chisq.test(cp_exng)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cp_exng
## X-squared = 67.348, df = 3, p-value = 1.577e-14
```

4.3.5. Modelo de regresión logística

En este apartado, se creará un modelo de regesión logística con el fin de determinar si es posible determinar la posibilidad de padecer un ataque cardíaco en función de los valores que tomen las variables independientes.

Con el fin de establecer la variable objetivo como variable dependiente, se transformarán los valores 1 y 2 de la variable a 0 y 1, respectivamente.

```
# Se realiza el reemplazo correspondiente.
df$output<- ifelse(df$output == 1, 0, 1)
# Se crea un archivo heart_def.csv con los datos finales analizados.
write.csv(df, "D:\\43591894v\\Downloads\\heart_def.csv")
```

Posteriormente, se crea un modelo de regresión logística con el conjunto de entrenamiento.

```
# Se crea el modelo de regresión logística teniendo en cuenta todas las variables como independientes.
modelo <- glm(formula = output ~ .,
              family = binomial,
              data = entrenamiento)
```

Tras obtener el resumen del modelo, se observan los siguientes aspectos:

- Las variables *sex1*, *caa1* y *caa2* son significativas a un nivel de confianza del 99,9%.
- Las variables *oldpeak* y *cp2* son significativas a un nivel de confianza del 99%.
- Las variables *cp3* y *caa3* son significativas a un nivel de confianza del 95%.

```
summary(modelo)
```

```
##
## Call:
## glm(formula = output ~ ., family = binomial, data = entrenamiento)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9831  -0.2595   0.1099   0.4338   2.8727
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.436e+01  1.455e+03   0.010 0.992125
## age          2.745e-02  2.818e-02   0.974 0.330104
## sex1        -1.625e+00  6.669e-01  -2.437 0.014803 *
## cp1          1.248e+00  6.831e-01   1.826 0.067794 .
## cp2          2.016e+00  6.279e-01   3.211 0.001324 **
## cp3          1.951e+00  8.019e-01   2.432 0.014997 *
## trtbps      -2.331e-02  1.345e-02  -1.734 0.082940 .
## chol        -4.722e-03  5.535e-03  -0.853 0.393596
```

```
## fbs1         3.930e-01  6.709e-01   0.586 0.558037
## restecg1     7.407e-01  4.583e-01   1.616 0.106077
## restecg2    -2.139e-01  3.320e+00  -0.064 0.948646
## thalachh     1.900e-02  1.459e-02   1.302 0.192852
## exng1       -9.528e-01  5.295e-01  -1.799 0.071940 .
## oldpeak     -7.155e-01  3.036e-01  -2.356 0.018453 *
## slp1        -1.407e+00  1.272e+00  -1.106 0.268755
## slp2        -4.499e-01  1.355e+00  -0.332 0.739801
## caa1        -1.916e+00  5.971e-01  -3.209 0.001331 **
## caa2        -3.122e+00  8.755e-01  -3.566 0.000362 ***
## caa3        -1.815e+00  1.059e+00  -1.714 0.086616 .
## caa4        -3.773e-01  2.965e+00  -0.127 0.898755
## thall1      -1.123e+01  1.455e+03  -0.008 0.993845
## thall2      -1.112e+01  1.455e+03  -0.008 0.993903
## thall3      -1.270e+01  1.455e+03  -0.009 0.993040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 335.05  on 242  degrees of freedom
## Residual deviance: 138.82  on 220  degrees of freedom
## AIC: 184.82
##
## Number of Fisher Scoring iterations: 14
```

A continuación, se predice con el conjunto de test, se evalúa la bondad del modelo mediante el *auc* y se obtienen los residuales del modelo.

```
# Se realizan predicciones en el conjunto de test.
predicciones <- predict(modelo, prueba, type = "response")
# Se calcula el porcentaje de predicciones correctas realizadas por el modelo.
aciertos <- mean(predicciones > 0.5 && as.logical(prueba$output))
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
# Se calcula la curva roc de las predicciones.
roc <- roc(prueba$output, predicciones)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```r
c_statistic <- as.numeric(roc$auc)
library(ROCR)
# Se calcula el auc.
pred <- prediction(predicciones, prueba$output)
perf <- performance(pred, "auc")
auc <- as.numeric(perf@y.values)
# Se calculan los residuos del modelo.
residuales <- resid(modelo, type = "deviance")
```

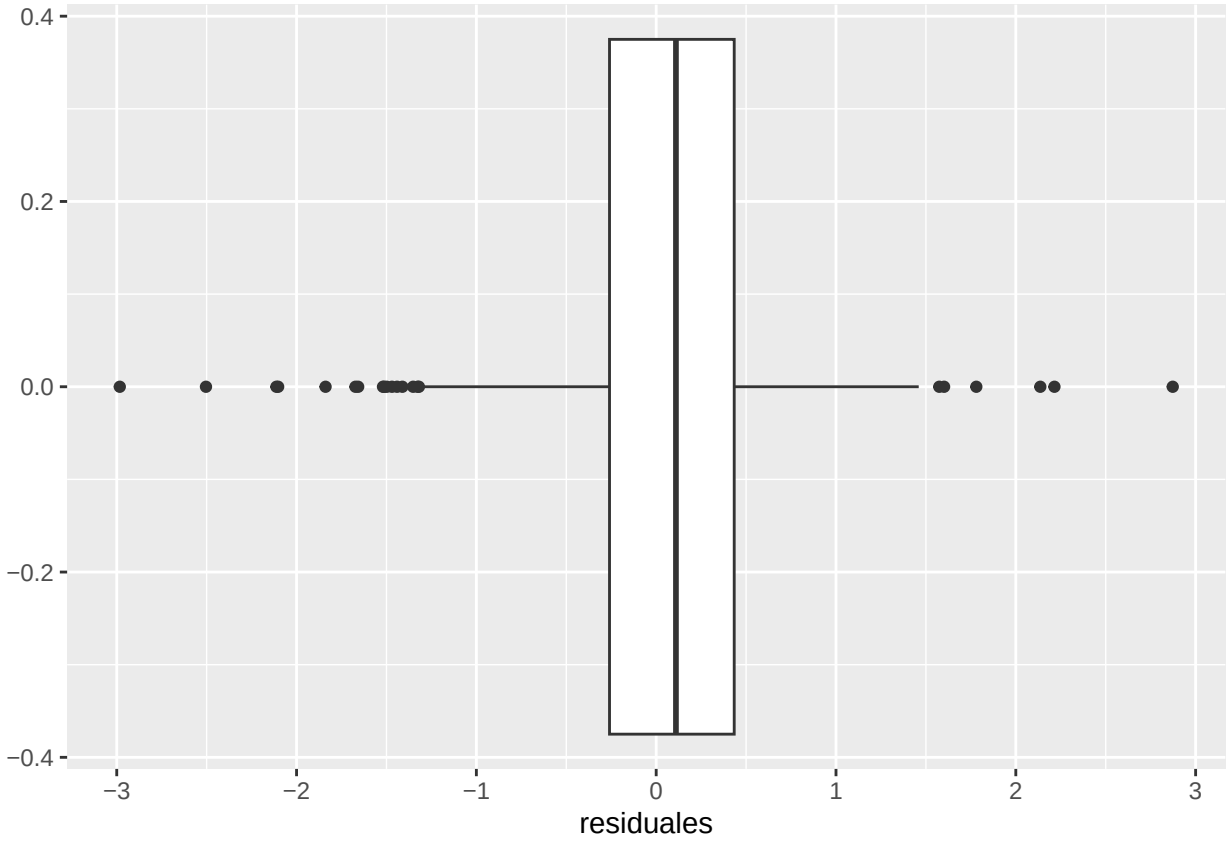
Se observa que el *auc* del modelo es satisfactorio:

```r
paste("El auc del modelo es", round(auc,2))
```

```
## [1] "El auc del modelo es 0.87"
```

Finalmente, se representan los residuales en un boxplot con el fin de evaluar las predicciones del modelo. En este caso la mediana se encuentra cerca de 0, lo que significa que la mayoría de las predicciones del modelo son bastante precisas. Sin embargo, se puede ver que existen algunos outliers lo cual indica que el modelo no está ajustando adecuadamente a algunas de las observaciones en el conjunto de datos. Pese a ello, se trata de un modelo que en este contexto es bastante satisfactorio.

```r
library(ggplot2)

ggplot(data = data.frame(residuales), aes(x = residuales)) +
  geom_boxplot()
```



4.3.6. Test de varianza ANOVA

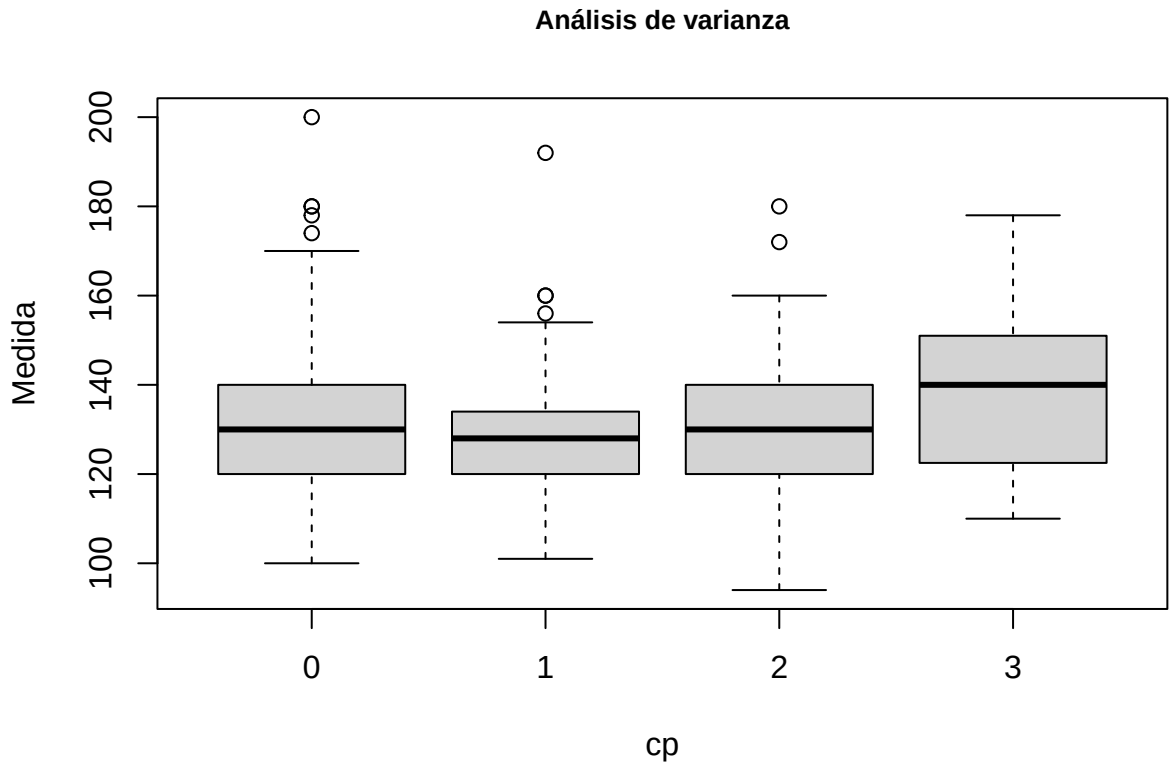
En este punto se plantea realizar un analisis de varianza ANOVA de un factor para determinar si existen diferencias significativas por lo que se refiere a la tensión arterial en función del tipo de dolor torácico. Este test considera que un p-valor menor de 0.05 indica una diferencia significativa entre los grupos o tratamientos. Por lo tanto, en este caso, con un p-valor de 0.0344, se podría concluir que hay diferencias significativas entre los tratamientos.

```
aov_modelo <- aov(formula = trtbps ~ cp, data = df)
summary(aov_modelo)
```

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

Finalmente, este hecho también puede observarse visualmente mediante un boxplot que contenga ambas variables.

```
boxplot(trtbps ~ cp, data = df, ylab = "Medida", main = "Análisis de varianza", cex.main=0.8)
```



5. Resolución del problema

Por último, en este punto se comentarán los resultados del trabajo realizado a lo largo de esta práctica.

- Se ha demostrado que las variables numéricas del conjunto de datos pueden aproximarse a una distribución normal según el Teorema del Límite Central.
- Se ha demostrado que existe homoscedasticidad entre los grupos de datos comparables (Las edades por sexo y la tensión arterial en función del tipo de dolor torácico) mediante técnicas que asi lo requieren como ANOVA y la prueba t de Student.
- El estudio de correlación lineal entre variables numéricas indica que no existen correlacion especialmente relevente entre las mismas. Sin embargo, se observan ligeras correlaciones entre la edad y algunas variables como el colesterol, la Depresión del ST inducida por el ejercicio, la frecuencia cardíaca o la tensión arterial.
- El estudio diferencias significativas entre las variables categóricas y la variable objetivo apunta a que existe una diferencia significativa entre las variables *sex*, *cp*, *restecg*, *exng*, *slp*, *caa*, *thall* y la variable objetivo. Por contra, este no es el caso de la variable *fbs* y la variable objetivo.Por otro lado, cabe comentar que el test en algunas variables arroja un mensaje indicando que la aproximación de la prueba puede no ser correcta. En este caso en concreto, se debe a que la prueba de chi cuadrado requiere que las frecuencias esperadas para cada categoría sean suficientemente grandes. Por lo tanto, como parte de un trabajo futuro seria adecuado probar otra prueba de hipótesis alternativa apropiada para los datos y los supuestos a evaluar.
- La prueba t de Student niega el hecho de que existan diferencias significativas por lo que se refiere a la edad en función del sexo.
- La prueba chi-cuadrado afirma que existen diferencias significativas entre las variables *exng* y *cp*.
- El modelo de regresión logística creado para precedir la posibilidad de que un paciente tenga un infarto cardíaco muestra resultados satisfactorios con un auc superior al 80%. Asimismo, revela que las variables significativas para predecir dicha posibilidad son: *sex1*, *caa1*, *caa2*, *oldpeak*, *cp2*, *cp3* y *caa3*.
- El test anova afirma que existen diferencias entre los grupos de la variable *trtbps* en función de *cp* son diferentes.

6. Recursos bibliográficos

[1] Kaggle. Heart Attack Analysis & Prediction Dataset. Obtenido de: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

[2] UCI. Machine Learning Repository. Heart Disease Data Set. Obtenido de: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3] Mayo Clinic. Crisis hipertensiva: ¿cuáles son los síntomas?. Obtenido de: https://www.mayoclinic.org/es-es/diseases-conditions/high-blood-pressure/expert-answers/hypertensive-crisis/faq-20058491

[4] ELO Smart Nurtition. TOTAL CHOLESTEROL OVERVIEW. Obtenido de: https://www.elo.health/biomarkers/total-cholesterol-overview/250/

[5] Bupa Salud. Taquicardia Supraventricular. Obtenido de: https://www.bupasalud.com/salud/taquicardia-supraventricular

[6] Revista Española de Cardiologia. Obtenido de https://www.revespcardiol.org/es-correlacion-angiografica-del-descenso-del-articulo-X0300893297004675

[7] Universitat Oberta de Catalunya. Introducción a la limpieza y análisis de los datos.