PRÁCTICA 2

Víctor Manuel Miñambres y Sua de la Cruz

9 de enero de 2024

Contents

Carga y resumen de los datos

En primer lugar, vamos a cargar y resumir los datos, con la finalidad de tener una priemra impresión de los datos.

```
ruta <- file.path(dirname(getwd()), "data", "heart.csv")
data <- read.csv(ruta)
head(data)
```

```
##   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(data)
```

```
##       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
```

Limpieza de los datos

En este apartado vamos a realizar la limpieza de datos. Para ello, analizaremos los elementos vacíos que hay en el conjunto de datos, así como los ceros y su significado.

Ceros y elementos vacíos

```r
# Verificar los valores nulos y si cada columna contiene ceros
cat("Los datos", ifelse(any(is.na(data)), "SÍ", "NO"), "contienen valores nulos\n")
```

```
## Los datos NO contienen valores nulos
```

```r
for (col in colnames(data)) {
  cat("La columna", col, ifelse(any(data[, col] == 0), "SÍ", "NO"), "contiene ceros\n")
}
```

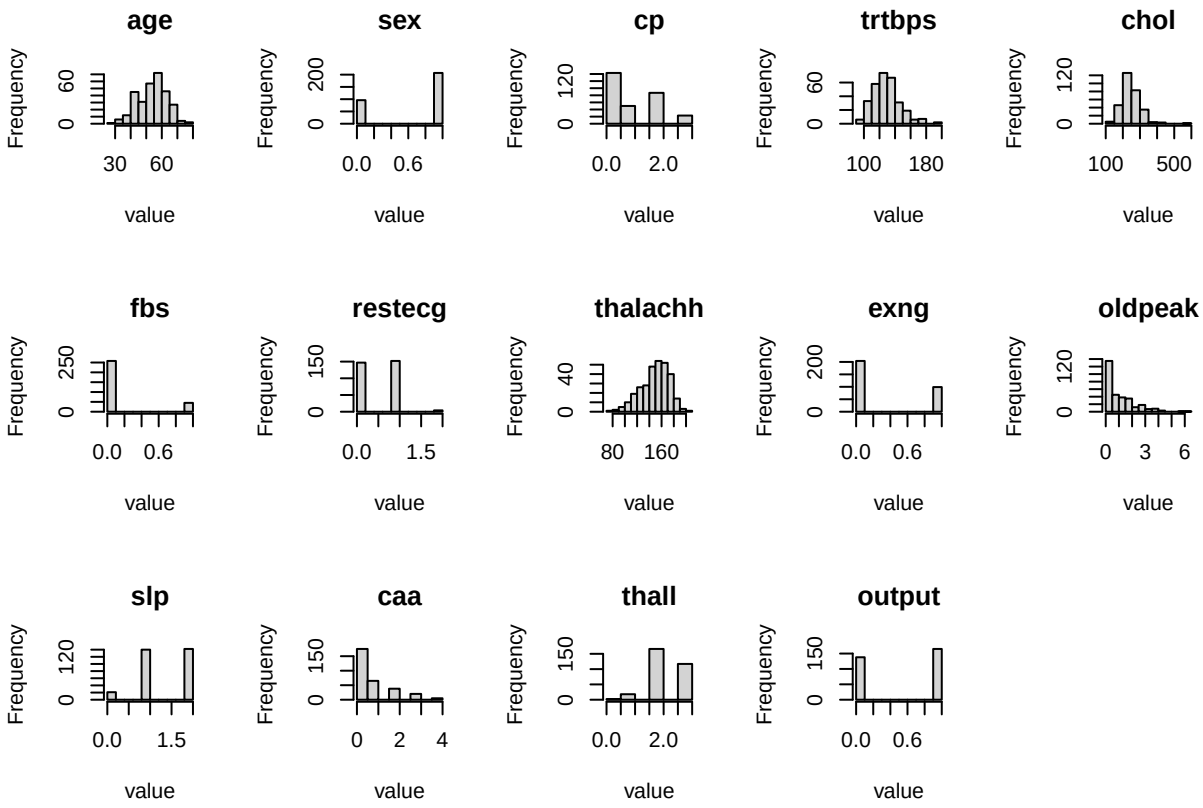
```
## La columna age NO contiene ceros
## La columna sex SÍ contiene ceros
## La columna cp SÍ contiene ceros
## La columna trtbps NO contiene ceros
## La columna chol NO contiene ceros
## La columna fbs SÍ contiene ceros
## La columna restecg SÍ contiene ceros
## La columna thalachh NO contiene ceros
## La columna exng SÍ contiene ceros
## La columna oldpeak SÍ contiene ceros
## La columna slp SÍ contiene ceros
## La columna caa SÍ contiene ceros
## La columna thall SÍ contiene ceros
## La columna output SÍ contiene ceros
```

En las columnas en las que encontramos ceros es coherente encontrar estos valores. Es decir, el número 0 es un indicador para algunas de las columnas (se interpreta como "NO" o "Falso" en alguno de los casos, en otros casos es un indicador que ha sido definido en la descripción del conjunto de datos). En el caso de la columna oldpeak, se representan la cantidad de milímetros (mm) de depresión del segmento ST en el electrocardiograma (ECG) durante la prueba de esfuerzo, por lo que también es coherente que haya valores de cero dependiendo de cómo se haya realizado la medición de los datos. Por ello, podemos concluir que se ha realizado una correcta limpieza sobre los datos, al menos respecto a los valores nulos y los ceros. A continuación vamos a analizar los valores extremos (outliers).

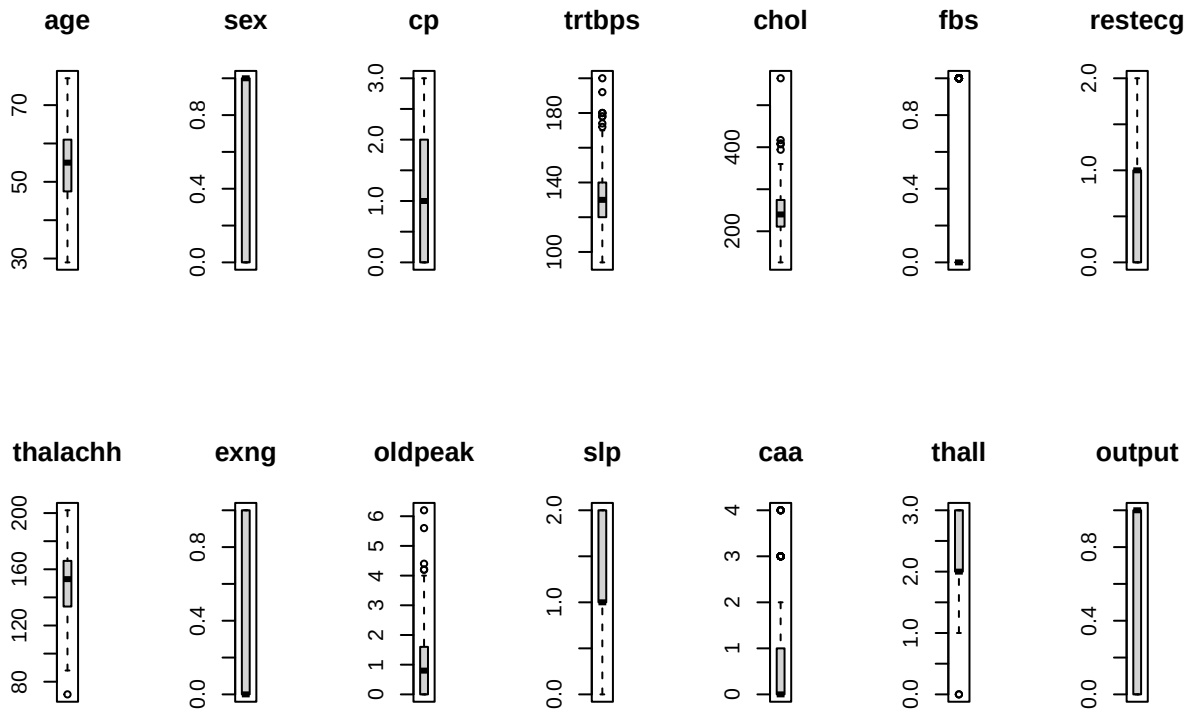
Valores extremos

En este apartado vamos a observar los valores extremos que hay en el conjunto de datos mediante histogramas y diagramas de caja.

```r
par(mfrow=c(3, 5))
for (col in colnames(data)) {
  hist(data[, col], main=col, xlab="value")
}
```




```
par(mfrow=c(2, 7))
for (col in colnames(data)) {
  boxplot(data[, col], main=col)
}
```

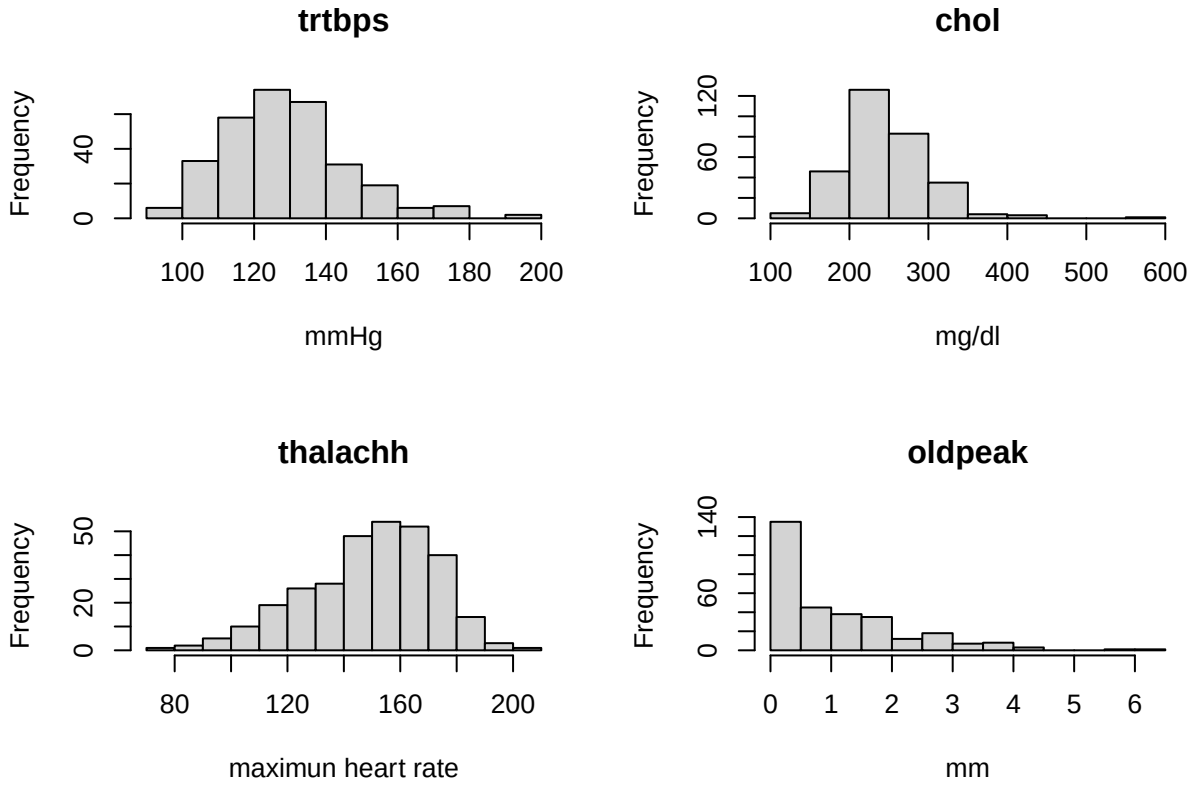


Como podemos observar, hay algunos casos en el que se detectan como valores extremos algunos de los códigos especificados para esa columna. Por ejemplo, en el caso de la columna fbs, sabemos que los códigos son 1 y 0, por lo que en este caso el valor 1 no es un valor extremo sino poco común. Lo mismo ocurre con la columna thall.

En el caso de la columna caa, ocurre algo parecido para el valor de 3, que se detecta como valor extremo cuando realmente es un código especificado. Sin embargo, para esta columna no se ha definido ningún código cuyo valor sea 4, por lo que en este caso sí podemos asumir que 4 es un valor erróneo. Como carecemos de contexto, no sabemos si este valor representa un código que desconocemos o si se debe a un error de codificación. En este caso, se puede optar por ignorar los casos en los que la columna caa tenga el valor 4 (cinco registros) o asumir que ha sido un error y cambiar el valor a 3. En nuestro caso, vamos a optar por ignorar estos registros.

Por otro lado, podemos observar que en las columnas trtbps, chol, thalachh y oldpeak se detectan valores extremos, por lo que vamos a hacer un análisis más detallado de estas columnas para determinar si son valores erróneos o no.

```
values <- c("mmHg", "mg/dl", "maximun heart rate", "mm")
names <- c("trtbps", "chol", "thalachh", "oldpeak")
par(mfrow=c(2, 2))
for (col in c(1:4)) {
  hist(data[, names[col]], main=names[col], xlab=values[col])
}
```



Teniendo en cuenta que el contexto del conjunto de datos son los ataques al corazón, no es raro que los valores de la columna trtbps (presión arterial en reposo) lleguen a 200. Por otro lado, respecto a la columna thalachh, una frecuencia cardíaca máxima de 71 latidos por minuto podría considerarse normal en reposo para muchos adultos, por lo que vamos a asumir que no es un valor erróneo, a pesar de ser extremo respecto al conjunto de datos.

Respecto a los valores extremos de las columnas chol y oldpeak, teniendo en cuenta el contexto del conjunto de datos, no son valores tan extremos como para considerarlos erróneos, por lo que no los vamos a modificar.

A continuación, ignoramos los registros en los que la columna caa tiene el valor de 4. Una vez realizada la limpeiza de datos, podemos comenzar con el análisis de datos.

```
data<-data[data$caa<4,]
```

Análisis de los datos

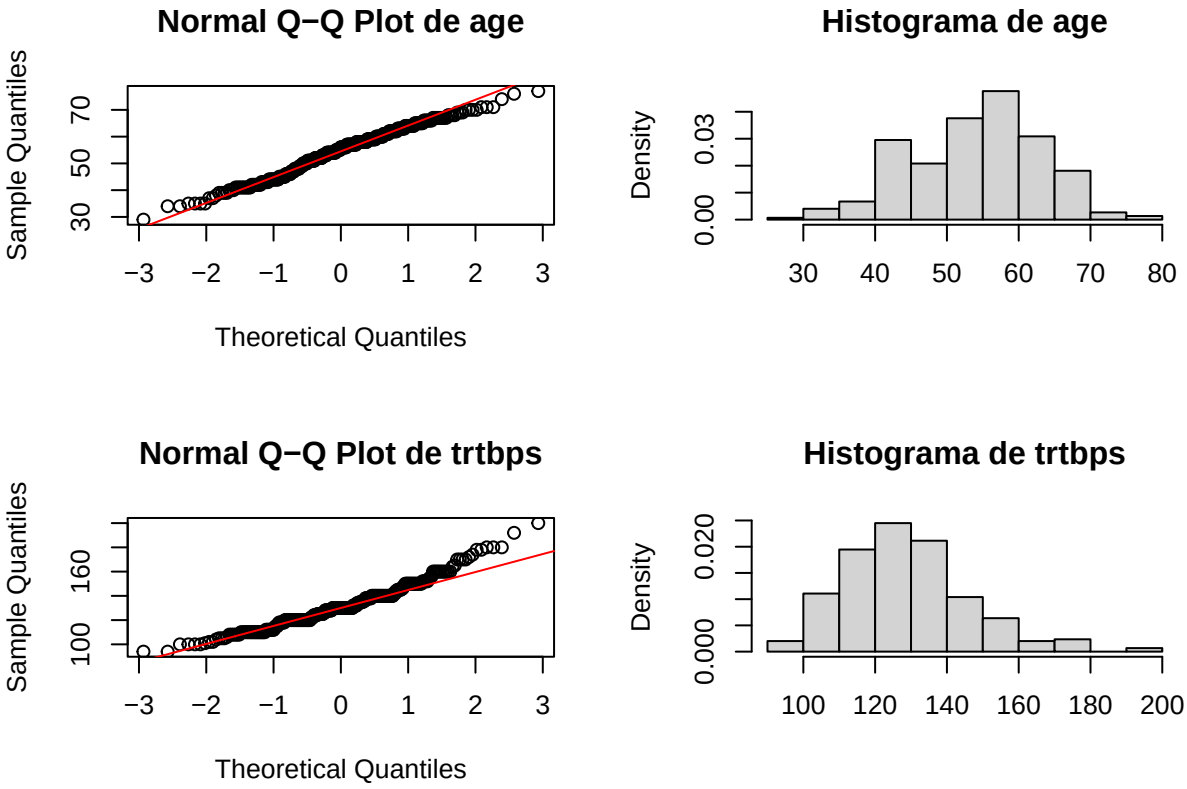
Por un lado, resulta interesante realizar un análisis comparativo entre pacientes con y sin ataques al corazón, con la finalidad de identificar las características clínicas de ambos grupos.

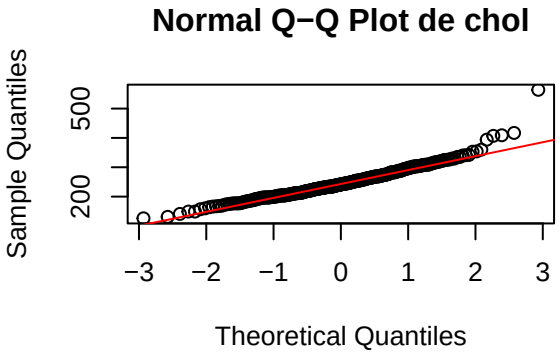
Por otro lado, podemos realizar otro análisis comparativo respecto al género, con el objetivo de explorar si hay diferencias significativas en las características clínicas entre hombres y mujeres, descubriendo qué

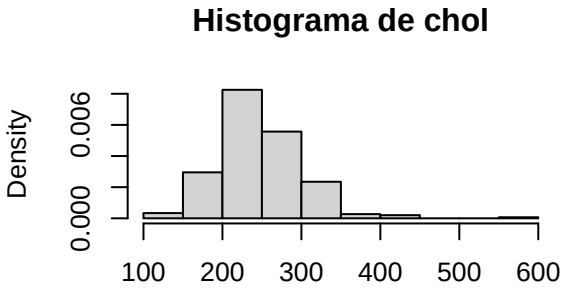
características cardíacas influyen más en la salud de cada género, como la frecuencia cardíaca máxima o la presión arterial.

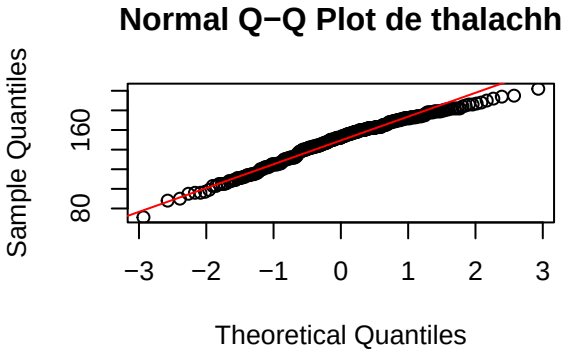
En primer lugar, vamos a comprobar la normalidad y homogeneidad de la varianza. Para ello, analizaremos la normalidad de aquellas variables que no sean categóricas. Para revisar si las variables pueden ser candidatas a la normalización miramos las gráficas de quantile-quantile plot y el histograma.

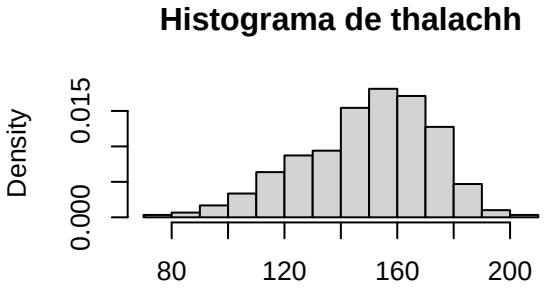
```r
# Convertir las columnas a tipo categórico
columnas_categoricas <- c("sex", "exng", "caa", "cp", "fbs", "restecg", "thall", "slp", "output")

par(mfrow=c(2,2))
for(i in colnames(data)) {
  if (!i%in%columnas_categoricas) {
    qqnorm(data[,i],main = paste("Normal Q-Q Plot de",i))
    qqline(data[,i],col="red")
    hist(data[,i],
      main=paste("Histograma de", i),
      xlab=colnames(data)[i], freq = FALSE)
  }
}
```

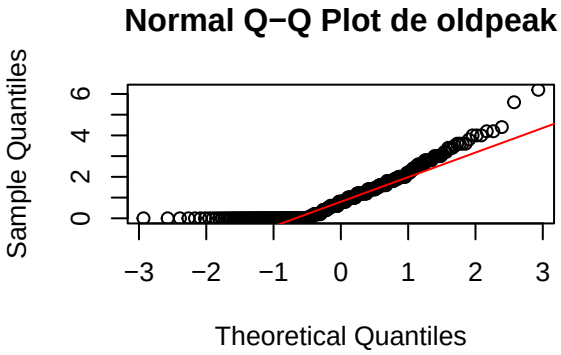


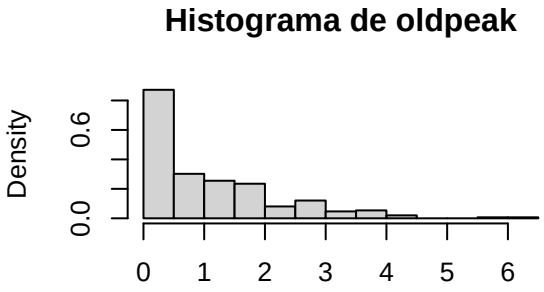

Normal Q−Q Plot de chol
Sample Quantiles
200
500
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histograma de chol
Density
0.000
0.006
100
200
300
400
500
600


Normal Q−Q Plot de thalachh
Sample Quantiles
80
160
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Histograma de thalachh
Density
0.000
0.015
80
120
160
200





Como podemos ver en los gráficos, las variables pueden ser candidatas a la normalización si es necesario, exceptuando la variable oldpeak. Aún así, observamos que en las columnas de la edad, trtbps y chol, los puntos en los extremos se desvían de la recta. Esto puede significar que en los extremos no siguen exactamente una distribución normal, siendo una posible causa la existencia de valores extremos.

Para revisar si las variables estan normalizadas vamos a aplicar el test de Shapiro Wilk en cada variable numérica.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data$age
## W = 0.98679, p-value = 0.007878
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data$trtbps
## W = 0.96558, p-value = 1.569e-06
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  data$chol
## W = 0.94696, p-value = 6.896e-09
```

```r
shapiro.test(data$thalachh)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$thalachh
## W = 0.97703, p-value = 0.0001028
```

```r
shapiro.test(data$oldpeak)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$oldpeak
## W = 0.84854, p-value < 2.2e-16
```

El p-value es inferior al coeficiente 0.05, por lo que se puede rechazar la hipotesis nula, significando que ninguna variable está normalizada. Aun así, según el teorema del límite central al tener más de 30 elementos en las observaciones podemos aproximarla como una distribución normal de media 0 y desviación estandard 1.

A continuación, vamos a comprobar la homogeneidad de la varianza mediante el test de levene:

```r
library(car)
```

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

```r
columnas_numéricas<-c("age", "trtbps", "chol", "thalachh", "oldpeak")
for (col in columnas_numéricas) {
  cat(col,":\n")
  print(leveneTest(data[[col]] ~ factor(data$output)))
  cat("\n")
}
```

```
## age :
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1  8.3708 0.004096 **
##       296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## trtbps :
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.7308 0.1893
##       296
##
## chol :
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1    0.14 0.7085
##       296
##
## thalachh :
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  5.3107 0.02189 *
##       296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## oldpeak :
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  30.553 7.134e-08 ***
##       296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos observar, para las variables trtbps y chol no hay evidencia significativa para rechazar la hipótesis nula, lo que significa que no se encuentra falta de homogeneidad de varianza para las variables. Sucede lo contrario con las variables age, thalachh y oldpeak, para las que sí hay evidencia para rechazar la hipótesis nula.

Aplicación de pruebas estadisticas

En primer lugar, vamos a generar un modelo de regresión lineal para intentar predecir si un paciente va a tener un ataque al corazón dependiendo de sus características clínicas.

```
modelo_lm <- lm(output ~ age + sex + cp + trtbps + chol + fbs + restecg + thalachh + exng + oldpeak + sl
summary(modelo_lm)
```

```
##
## Call:
## lm(formula = output ~ age + sex + cp + trtbps + chol + fbs +
##     restecg + thalachh + exng + oldpeak + slp + caa + thall,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93394 -0.19489  0.05217  0.25155  0.95521
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7769079  0.2896035   2.683 0.007732 **
## age          0.0008222  0.0027358   0.301 0.764005
## sex         -0.2019292  0.0464341  -4.349 1.91e-05 ***
## cp           0.1031040  0.0222314   4.638 5.38e-06 ***
## trtbps      -0.0023349  0.0012451  -1.875 0.061777 .
## chol        -0.0002432  0.0004165  -0.584 0.559778
## fbs          0.0400743  0.0595564   0.673 0.501570
## restecg      0.0348585  0.0395498   0.881 0.378857
## thalachh     0.0029058  0.0011130   2.611 0.009515 **
## exng        -0.1417676  0.0506919  -2.797 0.005516 **
## oldpeak     -0.0426322  0.0230830  -1.847 0.065801 .
## slp          0.0937086  0.0420642   2.228 0.026681 *
## caa         -0.1454863  0.0250666  -5.804 1.73e-08 ***
## thall       -0.1192956  0.0352644  -3.383 0.000818 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3485 on 284 degrees of freedom
## Multiple R-squared:  0.534,  Adjusted R-squared:  0.5126
## F-statistic: 25.03 on 13 and 284 DF,  p-value: < 2.2e-16
```

```
predicciones <- predict(modelo_lm, newdata = data)
umbral <- 0.5
predicciones_binarias <- ifelse(predicciones > umbral, 1, 0)
conf_matrix <- table(data$output, predicciones_binarias)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
cat("Precisión del modelo:", round(accuracy, 3), "\n")
```

```
## Precisión del modelo: 0.856
```

Podemos observar que las variables más significativas son el género, cp, caa y thall. Por ello, vamos a generar otros dos modelos, diferenciando por género, para ver el rendimiento de estos nuevos modelos y observar si las características clínicas relevantes cambian dependiendo del género.

```
modelo_hombres <- lm(output ~ age + cp + trtbps + chol + fbs + restecg + thalachh + exng + oldpeak + sl
modelo_mujeres <- lm(output ~ age + cp + trtbps + chol + fbs + restecg + thalachh + exng + oldpeak + sl
cat("Resumen del modelo para hombres:\n")
```

```
## Resumen del modelo para hombres:
```

```
summary(modelo_hombres)
```

```
##
## Call:
## lm(formula = output ~ age + cp + trtbps + chol + fbs + restecg +
##     thalachh + exng + oldpeak + slp + caa + thall, data = subset(data,
##     sex == 1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91075 -0.25992  0.07177  0.26375  0.92083
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.010e-01  3.907e-01   1.026 0.306037
## age          9.636e-05  3.638e-03   0.026 0.978895
## cp           1.002e-01  2.847e-02   3.519 0.000543 ***
## trtbps      -1.747e-03  1.709e-03  -1.022 0.308028
## chol        -3.747e-04  6.482e-04  -0.578 0.563891
## fbs          8.377e-02  7.685e-02   1.090 0.277049
## restecg      7.351e-02  5.365e-02   1.370 0.172278
## thalachh     3.688e-03  1.464e-03   2.519 0.012614 *
## exng        -7.324e-02  6.550e-02  -1.118 0.264913
## oldpeak     -5.345e-02  2.962e-02  -1.804 0.072793 .
## slp          8.144e-02  5.526e-02   1.474 0.142198
## caa         -1.455e-01  3.174e-02  -4.585 8.24e-06 ***
## thall       -1.062e-01  4.121e-02  -2.577 0.010739 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3754 on 189 degrees of freedom
## Multiple R-squared:  0.465,  Adjusted R-squared:  0.431
## F-statistic: 13.69 on 12 and 189 DF,  p-value: < 2.2e-16
```

```
predicciones <- predict(modelo_hombres, newdata = subset(data, sex == 1))
umbral <- 0.5
predicciones_binarias <- ifelse(predicciones > umbral, 1, 0)
conf_matrix <- table(subset(data, sex == 1)$output, predicciones_binarias)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
cat("Precisión del modelo:", round(accuracy, 3), "\n")
```

```
## Precisión del modelo: 0.832
```

```
cat("\nResumen del modelo para mujeres:\n")
```

```
##
## Resumen del modelo para mujeres:
```

```
summary(modelo_mujeres)
```

```
##
## Call:
## lm(formula = output ~ age + cp + trtbps + chol + fbs + restecg +
##     thalachh + exng + oldpeak + slp + caa + thall, data = subset(data,
##     sex == 0))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93891 -0.12498  0.03722  0.14136  0.55709
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.183e+00  4.349e-01   2.720  0.00794 **
```

```
## age          6.103e-04  4.056e-03   0.150  0.88076
## cp           1.008e-01  3.631e-02   2.775  0.00682 **
## trtbps      -1.796e-03  1.842e-03  -0.975  0.33226
## chol        -4.867e-05  5.408e-04  -0.090  0.92850
## fbs         -1.122e-01  1.006e-01  -1.115  0.26826
## restecg     -3.278e-02  5.753e-02  -0.570  0.57030
## thalachh     9.517e-04  1.787e-03   0.532  0.59585
## exng        -2.436e-01  8.619e-02  -2.827  0.00589 **
## oldpeak     -3.001e-02  3.840e-02  -0.782  0.43671
## slp          1.058e-01  7.082e-02   1.494  0.13904
## caa         -1.279e-01  4.405e-02  -2.904  0.00472 **
## thall       -2.021e-01  7.870e-02  -2.568  0.01201 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2862 on 83 degrees of freedom
## Multiple R-squared:  0.6223, Adjusted R-squared:  0.5677
## F-statistic:  11.4 on 12 and 83 DF,  p-value: 4.102e-13
```

```
predicciones <- predict(modelo_mujeres, newdata = subset(data, sex == 0))
umbral <- 0.5
predicciones_binarias <- ifelse(predicciones > umbral, 1, 0)
conf_matrix <- table(subset(data, sex == 0)$output, predicciones_binarias)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
cat("Precisión del modelo:", round(accuracy, 3), "\n")
```

```
## Precisión del modelo: 0.938
```

Analizando los resultados de ambos modelos, podemos observar que las características clínicas relevantes son parecidas para ambos géneros. De estas características destacan el cp (dolor de pecho) y el caa (cantidad de vasos sanguíneos principales que se pueden visualizar con fluoroscopia) como las más significativas. También podemos observar que si hacemos las predicciones por género, la precisión en las mujeres aumenta significativamente.

Por último vamos a analizar la correlación de las variables más significativas respecto al resto de variables.

```
library(corrplot)
```

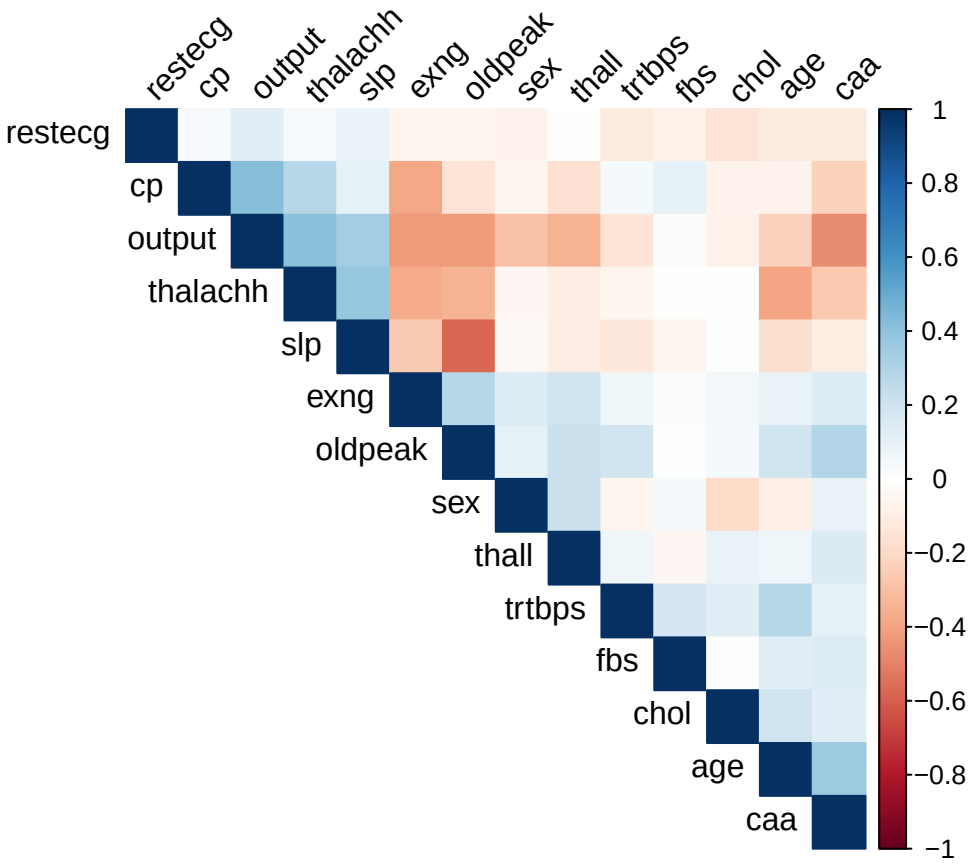
```
## corrplot 0.92 loaded
```

```
matriz_cor <- cor(data)
print(matriz_cor)
```

```
##                 age         sex          cp      trtbps         chol
## age      1.00000000 -0.08923879 -0.06089779  0.28891530  0.201646326
## sex     -0.08923879  1.00000000 -0.05472110 -0.05873146 -0.191568186
## cp      -0.06089779 -0.05472110  1.00000000  0.04440949 -0.067027335
## trtbps   0.28891530 -0.05873146  0.04440949  1.00000000  0.127441249
## chol     0.20164633 -0.19156819 -0.06702733  0.12744125  1.000000000
## fbs      0.12964064  0.04401701  0.10787602  0.18042482  0.008241061
## restecg -0.11228902 -0.06482406  0.03266716 -0.11710319 -0.141880559
## thalachh -0.39377454 -0.04983080  0.28855008 -0.05077097  0.000299409
```

```
## exng      0.09667105  0.14628850 -0.38788320  0.06969188  0.059177981
## oldpeak   0.19961944  0.10637096 -0.14443547  0.19714654  0.043976949
## slp      -0.16130884 -0.03478590  0.10942146 -0.12666364  0.005367318
## caa       0.36461047  0.09009946 -0.22528239  0.10283628  0.122054638
## thall     0.06592196  0.21039393 -0.15515072  0.06579541  0.096448389
## output   -0.22415051 -0.29011269  0.42715118 -0.14828288 -0.074976803
##                  fbs      restecg     thalachh        exng      oldpeak
## age       0.129640642 -0.112289018 -0.393774545  0.09667105  0.199619437
## sex       0.044017007 -0.064824060 -0.049830799  0.14628850  0.106370957
## cp        0.107876020  0.032667160  0.288550081 -0.38788320 -0.144435465
## trtbps    0.180424824 -0.117103187 -0.050770974  0.06969188  0.197146544
## chol      0.008241061 -0.141880559  0.000299409  0.05917798  0.043976949
## fbs       1.000000000 -0.072587091 -0.002689331  0.01067544  0.007844067
## restecg  -0.072587091  1.000000000  0.035678933 -0.05845991 -0.052390534
## thalachh -0.002689331  0.035678933  1.000000000 -0.37564956 -0.340350260
## exng      0.010675442 -0.058459905 -0.375649563  1.00000000  0.288083709
## oldpeak   0.007844067 -0.052390534 -0.340350260  0.28808371  1.000000000
## slp      -0.052514772  0.085200687  0.380474431 -0.25262492 -0.578529880
## caa       0.144643375 -0.110566127 -0.263376398  0.14412919  0.294309484
## thall    -0.042151168 -0.007129461 -0.090512679  0.20294003  0.213142358
## output   -0.014649116  0.124486843  0.417844406 -0.42919861 -0.429384028
##                  slp         caa        thall      output
## age      -0.161308844  0.36461047  0.065921962 -0.22415051
## sex      -0.034785898  0.09009946  0.210393931 -0.29011269
## cp        0.109421464 -0.22528239 -0.155150721  0.42715118
## trtbps   -0.126663635  0.10283628  0.065795412 -0.14828288
## chol      0.005367318  0.12205464  0.096448389 -0.07497680
## fbs      -0.052514772  0.14464338 -0.042151168 -0.01464912
## restecg   0.085200687 -0.11056613 -0.007129461  0.12448684
## thalachh  0.380474431 -0.26337640 -0.090512679  0.41784441
## exng     -0.252624916  0.14412919  0.202940033 -0.42919861
## oldpeak  -0.578529880  0.29430948  0.213142358 -0.42938403
## slp       1.000000000 -0.10803180 -0.096120778  0.34126697
## caa      -0.108031799  1.00000000  0.159395479 -0.46435731
## thall    -0.096120778  0.15939548  1.000000000 -0.34320334
## output    0.341266970 -0.46435731 -0.343203343  1.00000000
```

```
corrplot(matriz_cor, method = "color", type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```



En esta matriz de correlación podemos observar que la correlación más significativa para la variable cp es con la variable exng. Esta correlación puede tener sentido, ya que ambas variables representan el dolor en el pecho en diferentes situaciones. Por otro lado, la correlación más significativa para la variable caa es con el hecho de sufrir un ataque al corazón o no, lo que refuerza la idea de que la variable caa es realmente significativa a la hora de determinar si un paciente puede sufrir un ataque al corazón.

```
write.csv(data, file.path(dirname(getwd()), "data", "heart_out.csv"), row.names = FALSE)
```