PRÁCTICA 2: ¿Cómo realizar la limpieza y análisis de datos?

Lukaz Martin Doehne y Pablo Vadillo Berganza

08/01/2023

Índice

1. Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder?

Nuestro dataset está compuesto por 303 registros y 14 variables. Contiene información clínica de pacientes con el fin de poder predecir la probabilidad de fallo cardíaco.

Para ello, la variable a predecir, *output*, contiene información acerca del estrechamiento de los vasos sanguíneos obtenida a través de una angiografía. Si el estrechamiento es inferior al 50% toma el valor 0 y por tanto el paciente no se considera en riesgo de padecer una enfermedad del corazón. Por el contrario, si el estrechamiento es superior al 50%, la variable tomará el valor 1 y el paciente tendrá una mayor probabilidad de ataque al corazón.

A partir de aquí, será interesante realizar un análisis del resto de campos para ver cuáles son determinantes en la predicción del fallo cardíaco. Estudiar correlaciones, así como diferencias en la probabilidad del fallo cardíaco en diferentes grupos...

Pero antes, vamos a realizar una breve exploración del dataset.

```
## '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 ...
```

Todas las variables son numéricas.

```
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
```

Para terminar con este apartado, vamos a explicar brevemente todos y cada uno de los campos del dataset:

· **Age**: Edad de los pacientes en años. Toma valores entre 29 y 77. La media es 54,37.

· **Sex**: Sexo de los pacientes. (1 = hombre; 0 = mujer).

· **Cp**: Dolor de pecho. (1 = angina típica; 2 = angina atípica; 3 = dolor no anginal; 4 = asintomático).

· **Trtbps**: Presión arterial en reposo. Se trata del valor tomado en el ingreso al hospital, en mm Hg. Toma valores entre 94 y 200, siendo la media de 131,6.

· **Chol**: Colesterol sérico. Medido en mg/dl. Toma valores entre 126 y 564, con una media de 246,3.

· **Fbs**: Nivel de azúcar en sangre en ayunas. (1 = > 120 mg/dl; 0 = <= 120 mg/dl).

· **Restecg**: Resultados del electrocardiograma en reposo. (0 = normal; 1 = onda ST-T anómala; 2 = hipertrofia ventricular izquierda).

· **Thalachh**: Máximo pulso cardíaco obtenido. Toma valores entre 71 y 202. La media es 149,6.

· **Exng**: Angina inducida del ejercicio. (1 = sí; 0 = no).

· **Oldpeak**: Depresión del segmento ST inducida por ejercicio relativo al descanso. Toma valores entre 0 y 6,2, con una media de 1,04.

· **Slp**: Pendiente del segmento ST del pico del ejercicio. (1 = ascendente; 2 = plano; 3 = descendente).

· **Caa**: Número de los principales vasos sanguíneos coloreados por la fluoroscopia. Toma valores entre 0 y 4, con una media de 0,73.

· **Thall**: Talasemia. Menor nivel de hemoglobina. (1 = defecto fijo; 2 = normal; 3 = defecto reversible). El defecto fijo hace referencia a un defecto que ocurre tanto en reposo como durante el esfuerzo. El defecto reversible, por el contrario, hace referencia a un defecto que ocurre durante el esfuerzo que no existía durante el reposo.

· **Output**: La variable a predecir. Diagnóstico de fallo cardíaco (0 = Estrechamiento de vasos sanguíneos $< 50\%$; 1 = Estrechamiento de vasos sanguíneos $> 50\%$).

2. Integración y selección

En este apartado analizaremos qué campos son realmente significativos a la hora de predecir la variable output.

En primer lugar, calcularemos los coeficientes de correlación de Pearson, y nos fijaremos expresamente en la variable a predecir, output, con el resto de variables independientes.

```
cor_pearson <- cor(df)

# Mostramos los valores de output con el resto,

cor_pearson[14,]
```

```
##         age         sex          cp      trtbps        chol         fbs
## -0.22543872 -0.28093658  0.43379826 -0.14493113 -0.08523911 -0.02804576
##     restecg    thalachh        exng     oldpeak         slp         caa
##  0.13722950  0.42174093 -0.43675708 -0.43069600  0.34587708 -0.39172399
##       thall      output
## -0.34402927  1.00000000
```

Las variables *chol* (colesterol sérico) y *fbs* (nivel de azúcar en sangre en ayunas) presentan un coeficiente por debajo del 0.1 (en valor absoluto), por lo que podemos concluir que la correlación **directa** con la variable a predecir es muy baja. Otras variables como *cp* presentan una correlación directa alta.

En el dataset en concreto, todas las variables parecen de valor. Si por ejemplo nos interesase estudiar solo hombres, ya que son más propensos a un paro cardíaco, podríamos quitar del dataset las mujeres y borrar la columna *sex*. O si nos interesas analizar solo personas menores de 60 años, ya que por encima las personas tienen un sistema inmunológico más débil, también podríamos quitarlas del dataset.

Creamos una copia del dataset, por si posteriormente queremos realizar transformaciones sobre los datos

```
df_def <- df
```

3. Limpieza de los datos

3.1. ¿Los datos contienen ceros o elementos vacíos?

Ante elementos vacíos se pueden seguir principalmente dos estrategias: la **imputación** y la **eliminación**.

Para la primera, se trata de rellenar los valores nulos basandose en los valores no vacíos. Esto se realizar típicamente mediante *knn*, interpolación o cogiendo la media de una columna.

La segunda estrategía consiste en eliminar las filas con valores nulos. Esta se suele usar cuando se tiene un dataset grande y los valores nulos no suponen un gran porcentaje.

Podemos comprobar que no se encuentran elementos nulos en el csv con el siguiente código:

```
sapply(df, function(x) sum(is.na(x)))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

Dado que no encontramos valores nulos no aplicamos ninguna de las estrategias. Pero vamos a discretizar la columna *age* para mejor procesamiento.

Discretización

Para la variable edad, en lugar de tener el valor exacto crearemos grupos por decenas de edad.

E.g:

· 30-39 -> 3

· 40-49 -> 4

...

Vemos el resultado

```
##
##   2   3   4   5   6   7
##   1  15  72 125  80  10
```

3.2. Identifica y gestiona los valores extremos

Los valores extremos (*outliers*) los identificaremos como aquellos que se encuentran fuera del rango [Lo,Ho] donde:

· Lo = Q1 – (1.5 * IQR)

· Ho = Q3 + (1.5 * IQR)

y IQR = Q3 - Q1. También hay una fórmula que trata outliers los que sobrepasan la media +/- la desviación estándar, pero optaremos por la fórmula descrita, ya que también es la que usa *R* por defecto en la función boxplot.

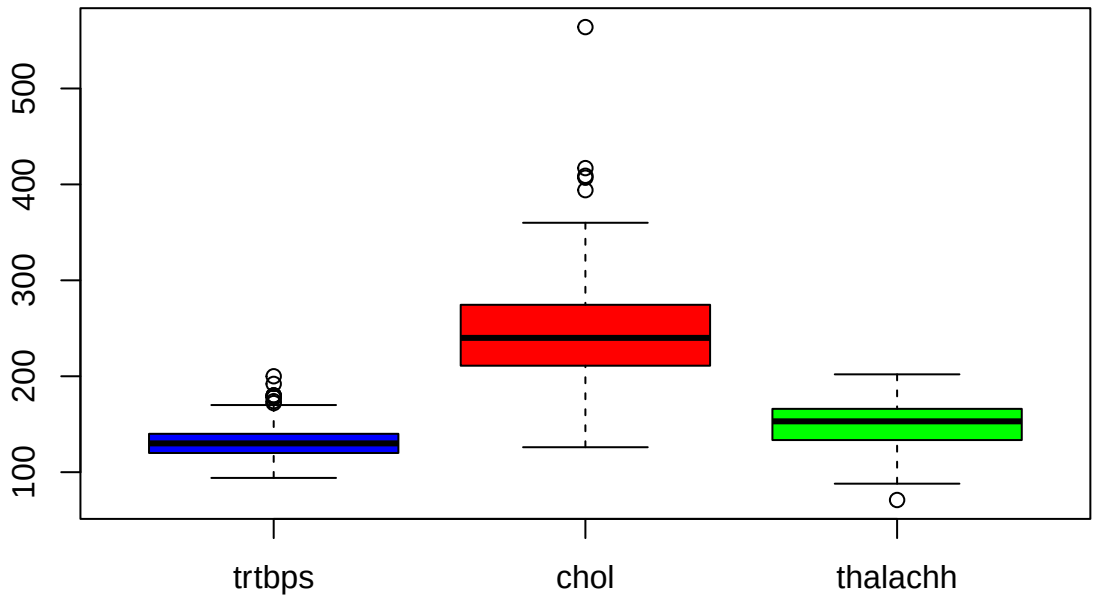
A continuación, mostramos la cantidad de outliers por columna con la siguiente función:

```r
outliers<-function(x){
  outliers<-boxplot.stats(x)$out
  return (length(outliers))
}
```

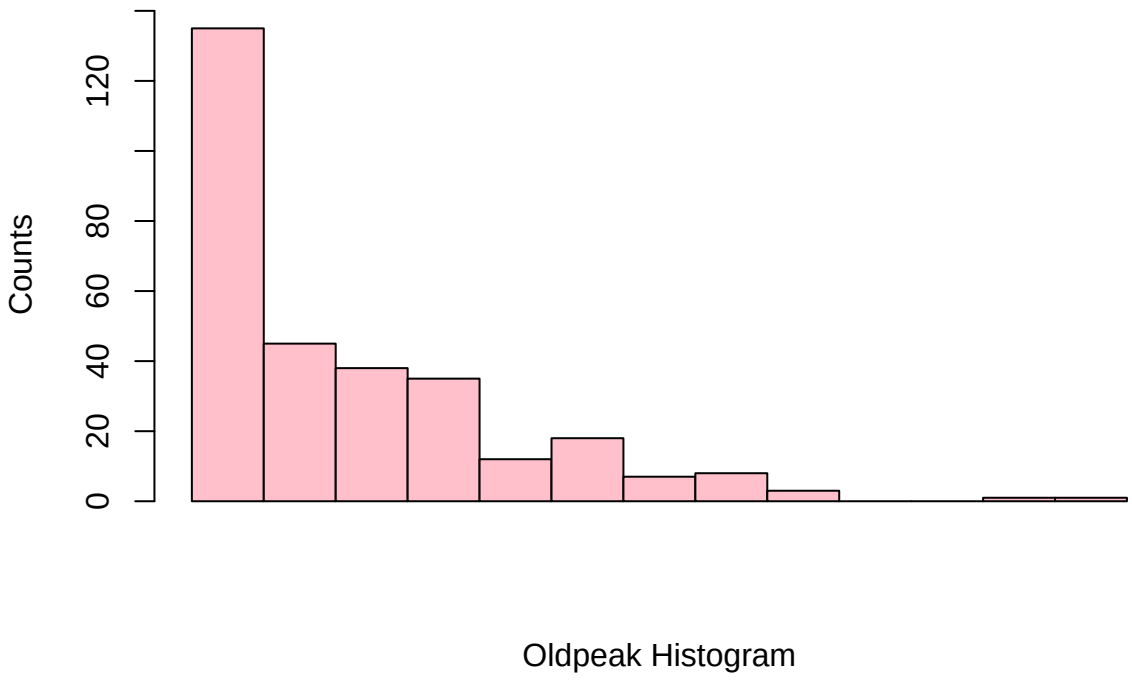
```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        9        5       45        0        1
##     exng  oldpeak      slp      caa    thall   output
##        0        5        0       25        2        0
```

Encontramos varios outliers en los datos. Las columnas *fbs*, *caa* y *thall* son columnas **númericas discretas** por lo que no los consideramos outliers, sino solo un dataset desequilibrado.

Visualizamos para entender mejor las demás variables.



La variable *oldpeak* la visualizamos en un histograma al estar en una escala diferente a las demás.



Viendo las visualizaciones suponemos que no ha habido errores a la hora de capturar los datos. Si que es cierto que en la columna *chol* encontramos un valor muy alejado del valor medio, pero lo mantendremos dentro del analisis al igual que el resto de los outliers. Si creamos un modelo de predicción podemos asegurar de incluir los outliers tanto en el train como en el test set.

4. Análisis de los datos

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

Hay 3 grupos en especial que nos interesa estudiar en función del fallo cardíaco:

· La variable **sex**. En función de hombres y mujeres.

· La variable **age**. En función de la edad de los pacientes.

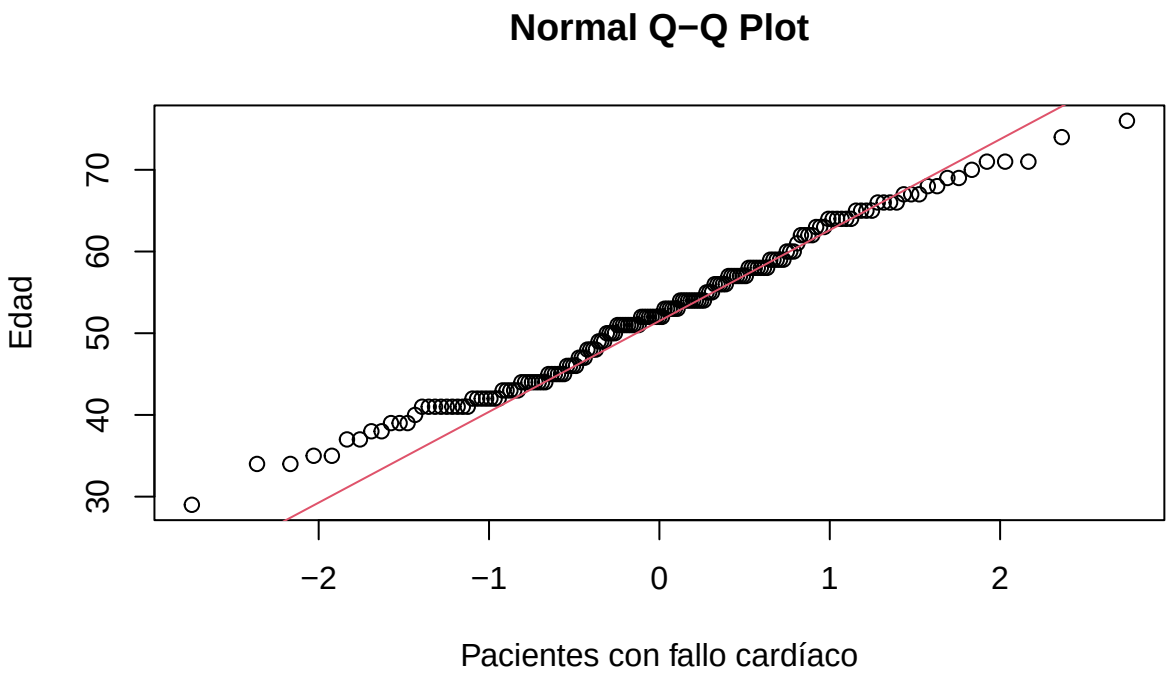
· La variable **trtbps**. En función de la presión arterial de los pacientes. Información extraída de: mayoclinic.org

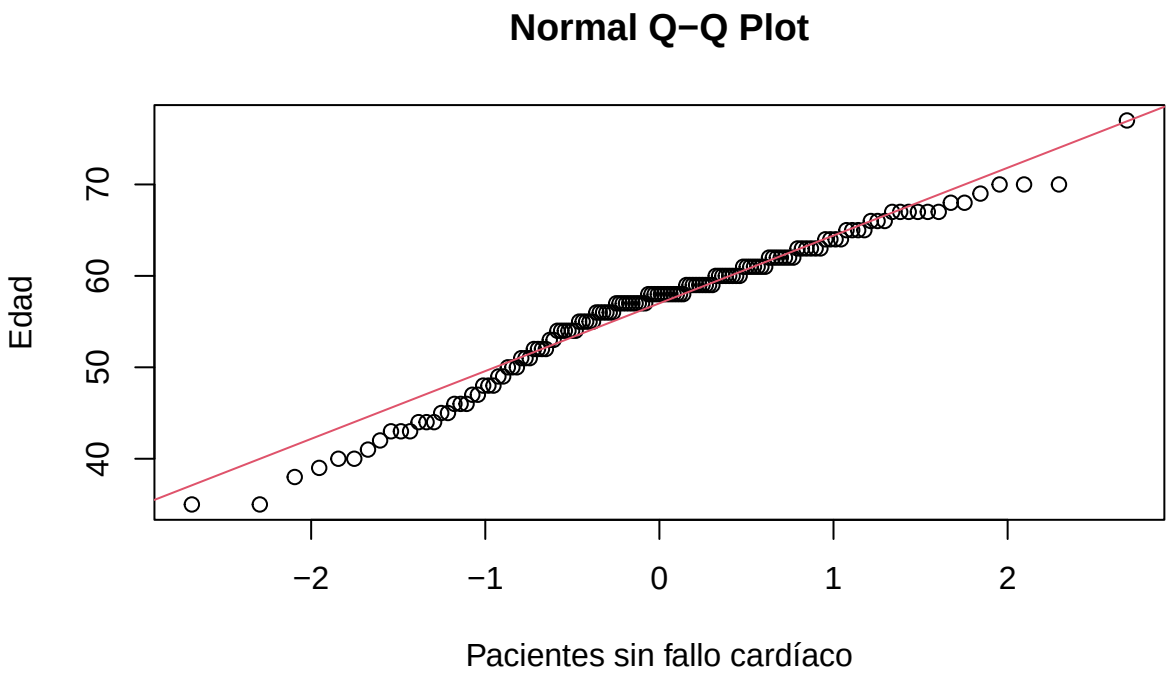
· La variable **output**. En función de la probabilidad de fallo cardíaco.

```
# Agrupación por género
df.male <- df_def[df_def$sex == 1,]
df.female <- df_def[df_def$sex == 0,]

# Agrupación por edad
df.young_adult <- df_def[df_def$age<40,]
df.adult <- df_def[df_def$age>=40&df_def$age<60,]
df.old_adult <- df_def[df_def$age>=60,]

# Agrupación por presión arterial
df.normal_pressure <- df_def[df_def$trtbps<120,]
df.high_pressure <- df_def[df_def$trtbps>=120&df_def$trtbps<130,]
df.hipertension1 <- df_def[df_def$trtbps>=130&df_def$trtbps<140,]
df.hipertension2 <- df_def[df_def$trtbps>=140&df_def$trtbps<180,]
df.crisis_hipertension <- df_def[df_def$trtbps>=180,]

# Agrupación por fallo cardíaco
df.high_output <- df_def[df_def$output == 1,]
df.low_output <-df_def[df_def$output == 0,]
```

4.2. Comprobación de la normalidad y homogeneidad de la varianza

Estudiamos la normalidad de la variable age y trtbps según la variable output, para ello, graficamos los QQplots que nos permiten observar similitudes con una distribución normal,





Visualmente, la variable age parece comportarse como una normal a partir de los gráficos, no obstante, realizaremos el test Shapiro para asegurarnos de que esto es cierto,

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  df.high_output$age
## W = 0.98677, p-value = 0.1211

##
##  Shapiro-Wilk normality test
##
## data:  df.low_output$age
## W = 0.96862, p-value = 0.002868
```

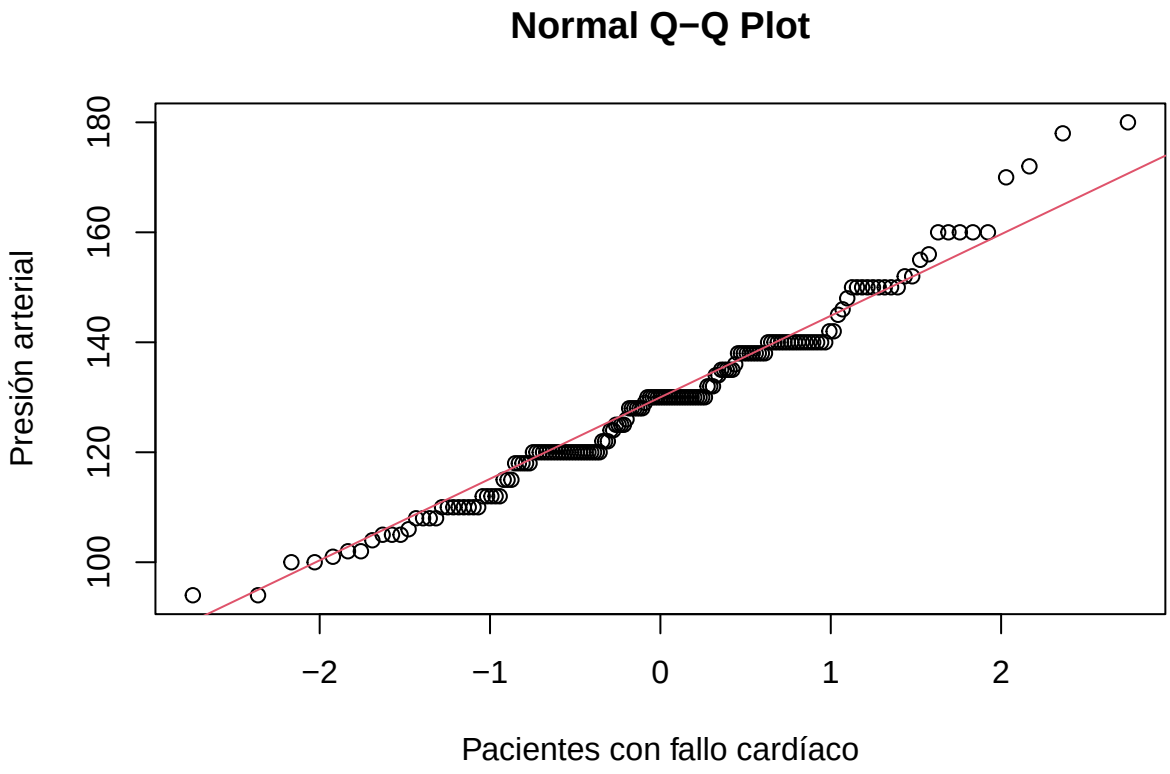
A partir del test de Shapiro-Wilk, podemos asumir que la variable edad para los pacientes con output = 1 sigue una distribución normal, ya que el p-valor es mayor que 0.05 (nivel de significancia normalmente utilizado), si bien para los pacientes con output = 0 esto no es cierto ya que el p-valor es menor.

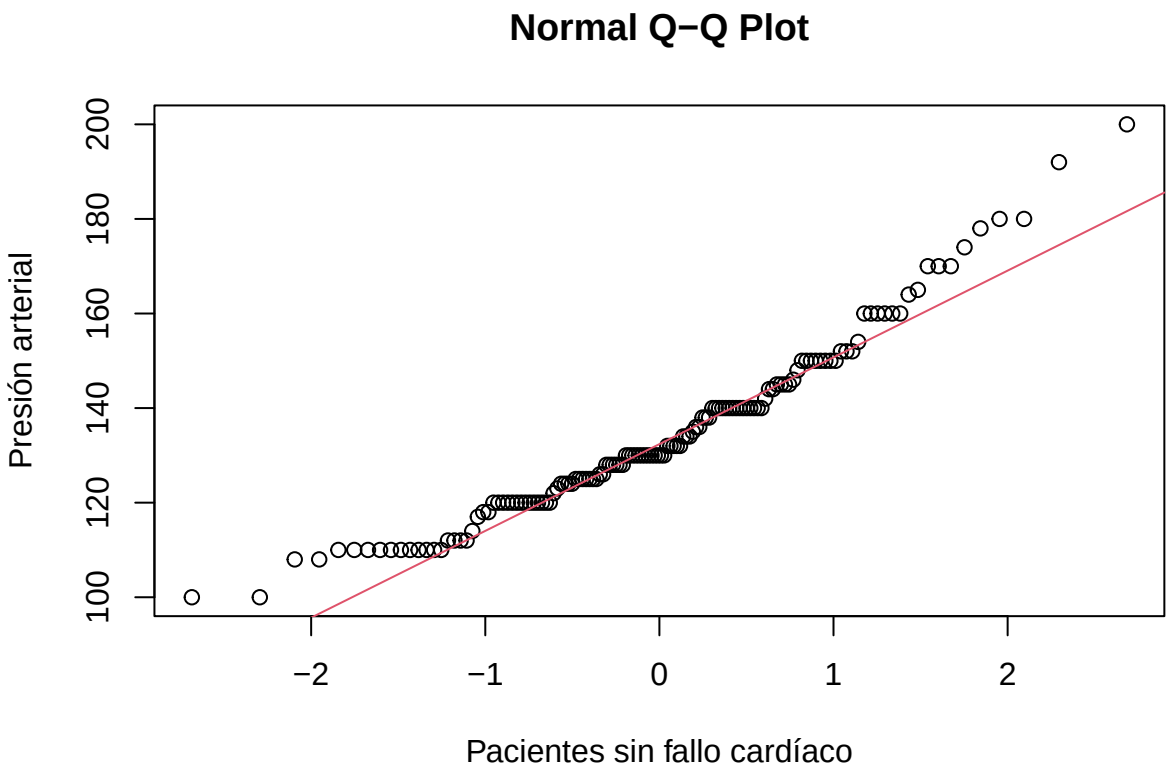
Como no podemos asumir normalidad para el conjunto de pacientes con output = 0, comprobamos igualdad de varianzas a partir del test fligner:

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  age by output
## Fligner-Killeen:med chi-squared = 7.2992, df = 1, p-value = 0.006898
```

De nuevo, tenemos un p-valor menor a 0.05, por lo que rechazamos la hipótesis nula de que las varianzas son iguales para ambos conjuntos.

Volvemos a hacer el mismo ejercicio, pero ahora para la variable trtbps,





De nuevo, tenemos que la variable trtbps parece comportarse como una normal a partir de los gráficos, no obstante, realizaremos el test Shapiro para asegurarnos de que esto es cierto,

```
##
##  Shapiro-Wilk normality test
##
## data:  df.high_output$trtbps
## W = 0.97865, p-value = 0.0119
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df.low_output$trtbps
## W = 0.95109, p-value = 8.365e-05
```

A partir del test de Shapiro-Wilk, tenemos que la variable trtbps no presenta una distribución normal en ninguno de los conjuntos, siendo sobre todo evidente en los pacientes con output = 0.

Como no podemos asumir normalidad, comprobamos igualdad de varianzas a partir del test fligner:

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  trtbps by output
## Fligner-Killeen:med chi-squared = 1.367, df = 1, p-value = 0.2423
```

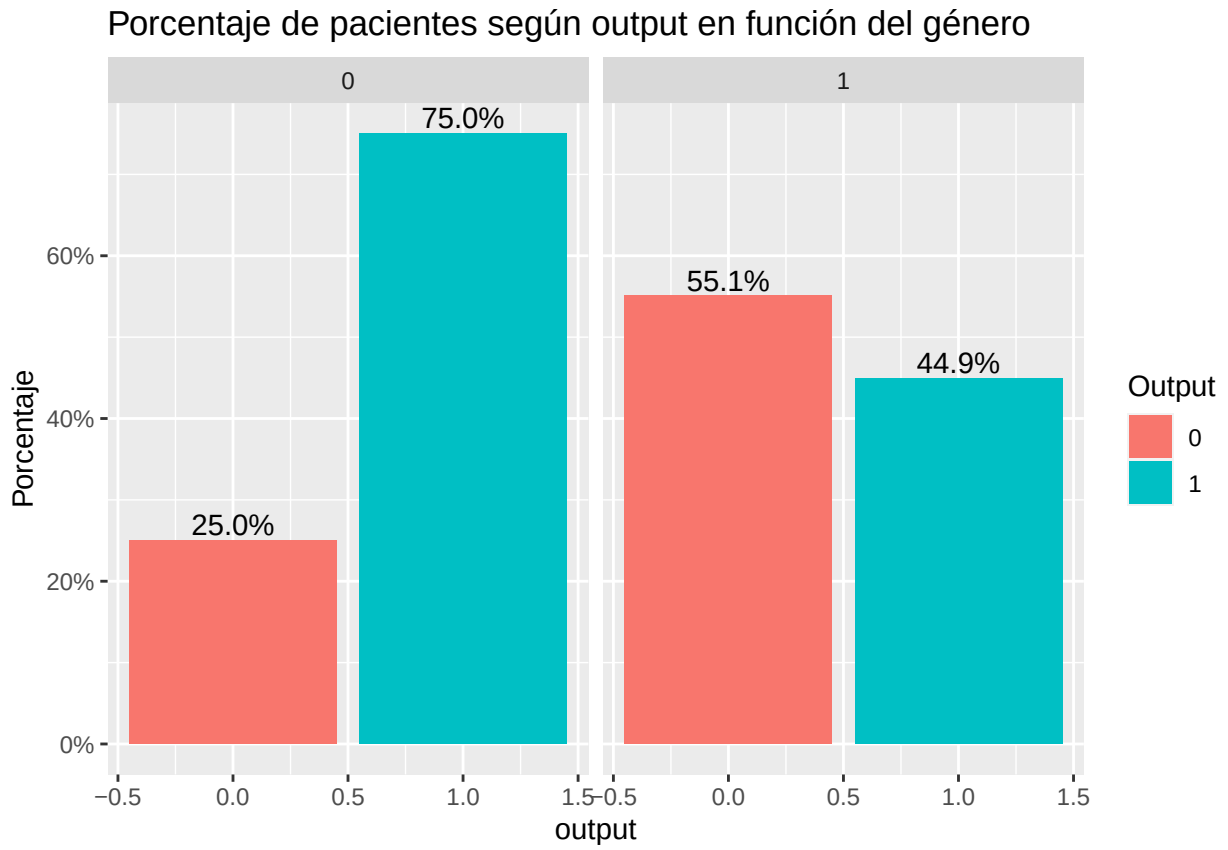
Obtenemos un p-valor de 0.243, por tanto no estamos en condiciones de rechazar la hipótesis nula y concluimos que las varianzas son similares para pacientes con output 1 y 0.

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

El primer test a realizar será un contraste sobre la proporción de una muestra. Queremos ver si hay diferencias en la proporción de pacientes con output = 1 entre hombres y mujeres.

Para ello, en primer lugar realizaremos un gráfico para ver visualmente cómo está distribuida nuestra muestra,

```
## Warning: The dot-dot notation (‘..prop..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(prop)‘ instead.
```



Según lo visualizado, entre las mujeres, el 75% presenta un output 1, mientras que entre los hombres sólo un 44,9%. Esto parece indicar que las mujeres presentan una mayor probabilidad de fallo cardíaco.

Por tanto, nuestra hipótesis nula será la igualdad de proporción de pacientes con output 1 entre hombres y mujeres, y la hipótesis alternativa que la proporción de pacientes con output 1 es mayor en mujeres que en hombres.

Aplicamos un contraste sobre la diferencia de proporciones, asumiendo la aproximación de la distribución binomial a una normal para muestras grandes. Se trata de un contraste unilateral por la derecha.

```
summary(factor(df.male$output))
```

```
##   0   1
## 114  93
```

```
summary(factor(df.female$output))
```

```
##  0  1
## 24 72
```

Dentro de las mujeres (96), (72) tienen una probabilidad alta de fallo cardíaco.

Entre los hombres (207), (93) tienen una probabilidad alta de fallo cardíaco.

Aplicamos el test,

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(72, 93) out of c(96, 207)
## X-squared = 23.914, df = 1, p-value = 5.036e-07
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2084305 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7500000 0.4492754
```

El p-valor es muy pequeño, por lo que rechazamos la hipótesis nula de igualdad de proporciones, y aceptamos la hipótesis de que las mujeres tienen un mayor probabilidad de fallo cardíaco.

A continuación, vamos a realizar dos tests de dos muestras independientes (según output) para descubrir si hay diferencias entre las variables age y trtbps.

En el apartado anterior ya vimos que no podíamos asumir normalidad en ninguno de los casos, por lo que aplicamos un test no paramétrico como el de Wilcoxon,

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df_def$age by df_def$output
## W = 14530, p-value = 3.439e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df_def$trtbps by df_def$output
## W = 12986, p-value = 0.03465
## alternative hypothesis: true location shift is not equal to 0
```

En ambos casos tenemos que el p-valor es menor a 0.05, por lo que rechazamos la hipótesis nula y asumimos que existen diferencias significativas en la edad y en la presión arterial de los pacientes según su probabilidad de fallo cardíaco.

Finalmente, podemos generar un modelo de regresión logística para predecir el valor de la variable output.

Para ello, dividimos el dataset en un conjunto de entrenamiento y otro de test:

```
set.seed(1)

sample <- sample(c(TRUE, FALSE), nrow(df_def), replace=TRUE, prob=c(0.8,0.2))
df.train  <- df_def[sample, ]
df.test   <- df_def[!sample, ]
```

Y creamos el modelo a partir del conjunto de entrenamiento,

```
Modlg <- glm(output ~ age + sex + cp + trtbps + thalachh + exng + oldpeak + slp + caa + thall, data = d
summary(Modlg)
```

```
##
## Call:
## glm(formula = output ~ age + sex + cp + trtbps + thalachh + exng +
##     oldpeak + slp + caa + thall, family = binomial, data = df.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5680  -0.3658   0.1530   0.5457   2.6701
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.12133    2.65584   1.175 0.239887
## age         -0.01517    0.02487  -0.610 0.541783
## sex         -1.57214    0.49190  -3.196 0.001393 **
## cp           0.94142    0.21840   4.310 1.63e-05 ***
## trtbps      -0.02363    0.01177  -2.008 0.044634 *
## thalachh     0.02334    0.01141   2.046 0.040747 *
## exng        -1.04691    0.47167  -2.220 0.026448 *
## oldpeak     -0.49048    0.23575  -2.081 0.037477 *
## slp          0.45558    0.41588   1.095 0.273313
## caa         -0.80186    0.20967  -3.824 0.000131 ***
## thall       -0.68279    0.32263  -2.116 0.034316 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 349.85  on 253  degrees of freedom
## Residual deviance: 173.66  on 243  degrees of freedom
## AIC: 195.66
##
## Number of Fisher Scoring iterations: 6
```

A continuación, generaremos la matriz de confusión para ver qué tal predice nuestro modelo, para ello haremos uso del conjunto de test.

```
prediction1 <- data.frame(predict(Modlg, df.test[,1:13], type = "response"))
prediction2 <- data.frame(ifelse(prediction1 < 0.5, 0, 1))

confusionMatrix(data=as.factor(prediction2$predict.Modlg..df.test...1.13...type....response..), referen
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 15  5
##          1  8 21
##
##                Accuracy : 0.7347
##                  95% CI : (0.5892, 0.8505)
##     No Information Rate : 0.5306
##     P-Value [Acc > NIR] : 0.00279
##
##                   Kappa : 0.4634
##
##  Mcnemar's Test P-Value : 0.57910
##
##             Sensitivity : 0.6522
##             Specificity : 0.8077
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.7241
##              Prevalence : 0.4694
##          Detection Rate : 0.3061
##    Detection Prevalence : 0.4082
##       Balanced Accuracy : 0.7299
##
##        'Positive' Class : 0
##
```

Nuestro modelo tiene un porcentaje de acierto del 73,47%. Si hablamos de predecir a los pacientes con mayor probabilidad de infarto (output = 1), esta probabilidad sube hasta el 80,77%.

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

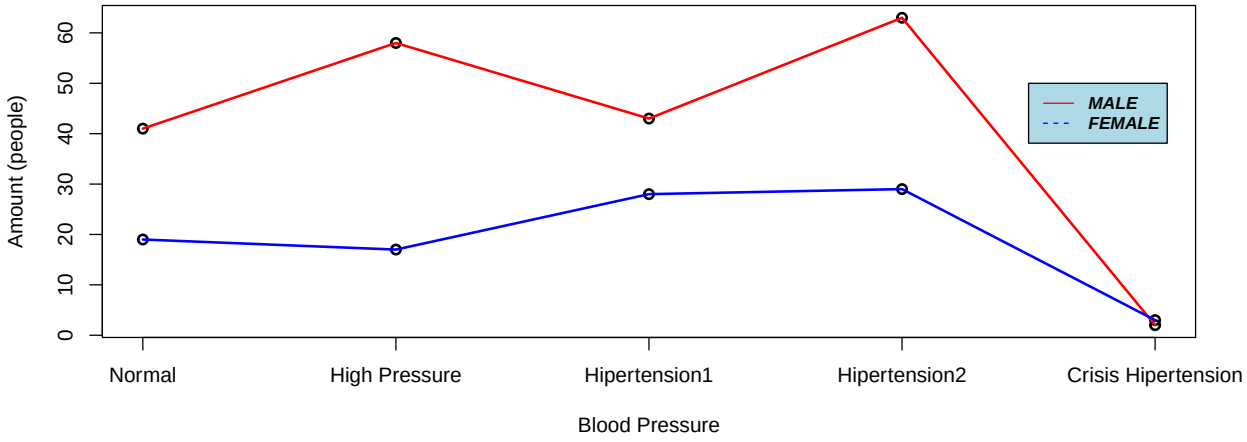
Primero, presentamos una tabla de la variable *sex* en función de *age* y el porcentaje de paros cardíacos.

sex	female	male
adult_under_40	100 %	63.64 %
adult_40_to_50	94.74 %	60.38 %
adult_50_to_60	70.59 %	45.05 %
adult_60_to_70	60.61 %	25.53 %
adult_over_70	100 %	20 %

Si miramos la fila *adult_40_to_50* de la tabla, vemos que la columna **female** tiene 94.74%. Indicando que de todas las mujeres de entre 40 y 50 del dataset, un 94.74% han tenido un paro cardíaco.

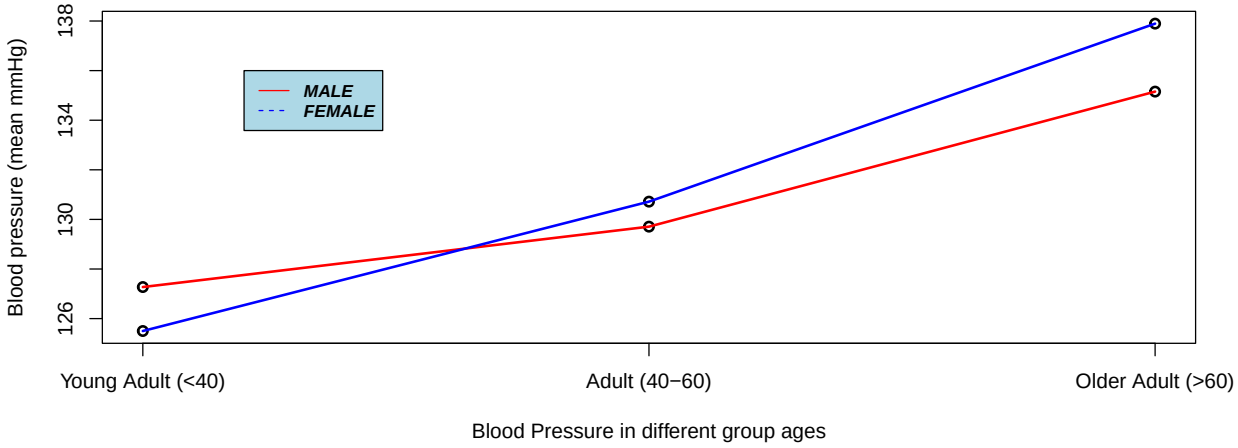
Concluimos pues que el porcentaje de mujeres con fallo cardíaco en función de mujeres totales tiene un ratio significativamente superior al de los hombres

También, vamos a hacer un lineplot de la variable *trtbps* en función de *sex*.



Podemos observar que la cantidad de hombres con niveles de hipertensión 2 y presión alta son superiores a los demás niveles. Para las mujeres parece ser más constante.

Finalmente, un lineplot de la variable *age* en función de la media de *trtbps*.



Aquí claramente también se aprecia una tendencia. A mayor edad, mayor presión arterial (tanto para hombre como para mujeres)

6. Resolución del problema