# Tipología y ciclo de vida de los datos: Práctica 2

Fanfan Yang

Enero 2022

Cargamos las librerías necesarias para la realización de esta PEC:

```
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
```

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
```

```
## Loading required package: ggplot2
```

```
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if(!require(grid)) install.packages('grid',repos='http://cran.us.r-
project.org'); library(grid)
```

```
## Loading required package: grid
```

```
if(!require(gridExtra)) install.packages('gridExtra',
repos='http://cran.us.r-project.org'); library(gridExtra)
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

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

```
if(!require(BSDA)){
    install.packages('ggpubr',repos='https://CRAN.R-
project.org/package=BSDA')
    library(BSDA)
}
```

```
## Loading required package: BSDA
```

```
## Warning: package 'BSDA' was built under R version 4.1.3

## Loading required package: lattice

##
## Attaching package: 'BSDA'

## The following object is masked from 'package:datasets':
##
##     Orange

if(!require(kableExtra)){
  install.packages('kableExtra',repos='https://CRAN.R-
project.org/package=kableExtra')
  library(kableExtra)
}

## Loading required package: kableExtra

## Warning: package 'kableExtra' was built under R version 4.1.3

##
## Attaching package: 'kableExtra'

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

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

El dataset se ha obtenido de Kaggle mediante el siguiente enlace: https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-predictiondataset

Contiene 303 registros de los pacientes con sus 14 caracteristicas con el objetivo de representar la relación del infarto con los distintos factores como edad, género, angina inducida por el ejercicio, número de vasos sanguíneos principales, tipo de dolor de pecho, presión arterial en reposo, etc., con el fin de predecir el riesgo de infarto de un paciente cualquiera teniendo los valores de estos factores.

Las 14 variables son:

- Age: edad del paciente, variable cuantitativa
- Sex: género (0 femenino y 1 masculino), variable binaria
- cp: tipo de dolor de pecho, variable categórica
  - Valor 0: angina típica
  - Valor 1: angina atípica
  - Valor 2: dolor no anginoso

  - Valor 3: asintomático
- trtbps: presión arterial en reposo (en mm Hg), variable cuantitativa
- chol: colesterol en mg/dl obtenido a través del sensor BMI, variable cuantitativa
- fbs: (Azúcar en sangre en ayunas > 120 mg/dl), variable binaria
  - 1 = true
  - 0 = false
- restecg: resultados electrocardiográficos en reposo, variable categórica
  - Valor 0: normal
  - Valor 1: tener anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0,05 mV)
  - Valor 2: hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes
- thalachh: frecuencia cardíaca máxima alcanzada, variable cuantitativa
- exng: angina inducida por el ejercicio (1: sí, 0: no), variable binaria
- old peak: depresión del ST inducida por el ejercicio en relación con el reposo, variable cuantitativa
- slp: la pendiente del segmento ST de ejercicio máximo, variable categórica
  - 0 = sin pendiente
  - 1 = plana
  - 2 = descendente
- caa: número de vasos sanguíneos principales, variable cuantitativa
- thall : talasemia, variable categórica
  - 0 = nula
  - 1 = defecto fijo
  - 2 = normal
  - 3 = defecto reversible
- output: variable binaria, diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfico)
  - 0: < 50% de estrechamiento del diámetro. Menos posibilidades de enfermedades del corazón
  - 1: > 50% de estrechamiento del diámetro. Más posibilidades de enfermedades del corazón

```
dat<-read.csv("./heart.csv", header=T,sep=",", stringsAsFactors = FALSE)
attach(dat)

str(dat)

## '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(dat)

##       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
```

Es importante para prevenir el infarto en los pacientes que son clasificados como altamente peligrosos (donde el valor objetivo es igual a 1) y también sirve como una guía para las personas sanas para identificar esa enfermedad mediante las síntomas presentadas.

## 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.

El dataset es relativamente pequeño, que solo contiene 303 registros y cada una de las variables presenta un factor que considera importante para la predicción de la enfermedad de corazón, por lo tanto, investigamos el dataset entero para encontrar las relaciones con cada uno de los factores.

Sin embargo, podemos mejorar el dataset modificando algunas variables como, por ejemplo, la variable sex, podemos sustituir sus valores (0 y 1) por (female y male) y la factorizamos para una mejor visualización.

```
dat["sex"][dat["sex"] == 0] <- 'female'
dat["sex"][dat["sex"] == 1] <- 'male'
dat$sex <- as.factor(dat$sex)
```

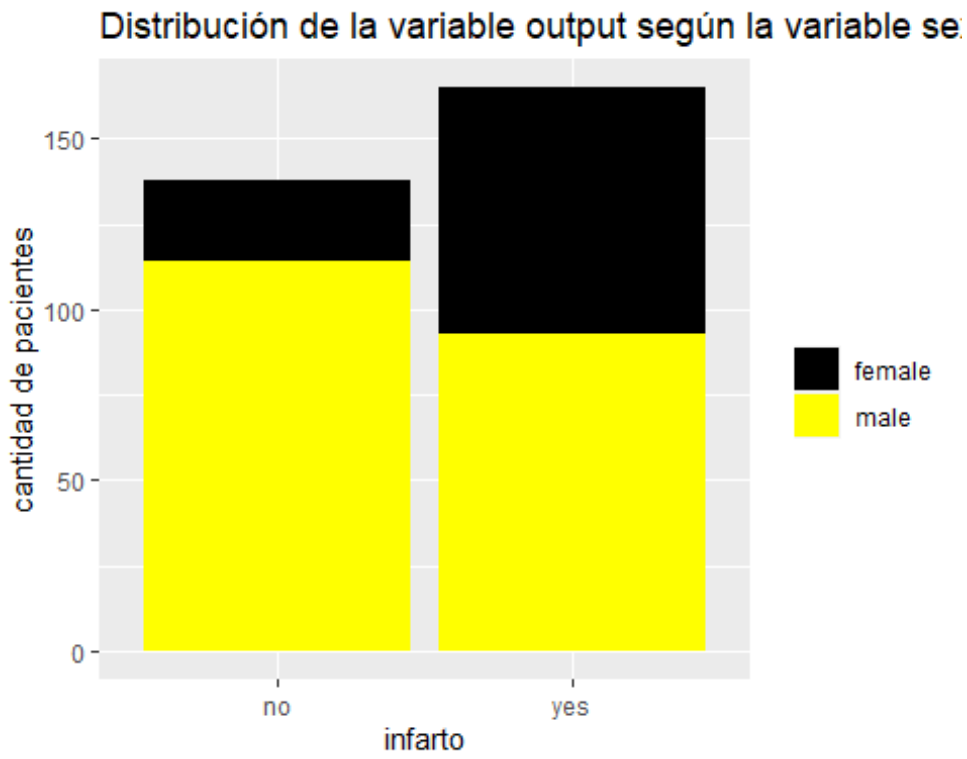
Y la variable objetivo output, podemos sustituir sus valores por yes y no en lugar de 1 y 0, donde yes significa la probabilidad alta del infarto (heart attack) y no la probabilidad baja del infarto.

```
dat["output"][dat["output"] == 0] <- 'no'
dat["output"][dat["output"] == 1] <- 'yes'
dat$output <- as.factor(dat$output)
```

Podemos mostrar la situación del infarto en los pacientes masculinos y las pacientes femeninas:

```
outputbySex<-ggplot(dat, aes(output, fill=sex))+geom_bar() +labs(x="infarto",
y="cantidad de pacientes")+ guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","yellow"))+ggtitle("Distribución de la
variable output según la variable sex")

grid.arrange(outputbySex, ncol=1)
```

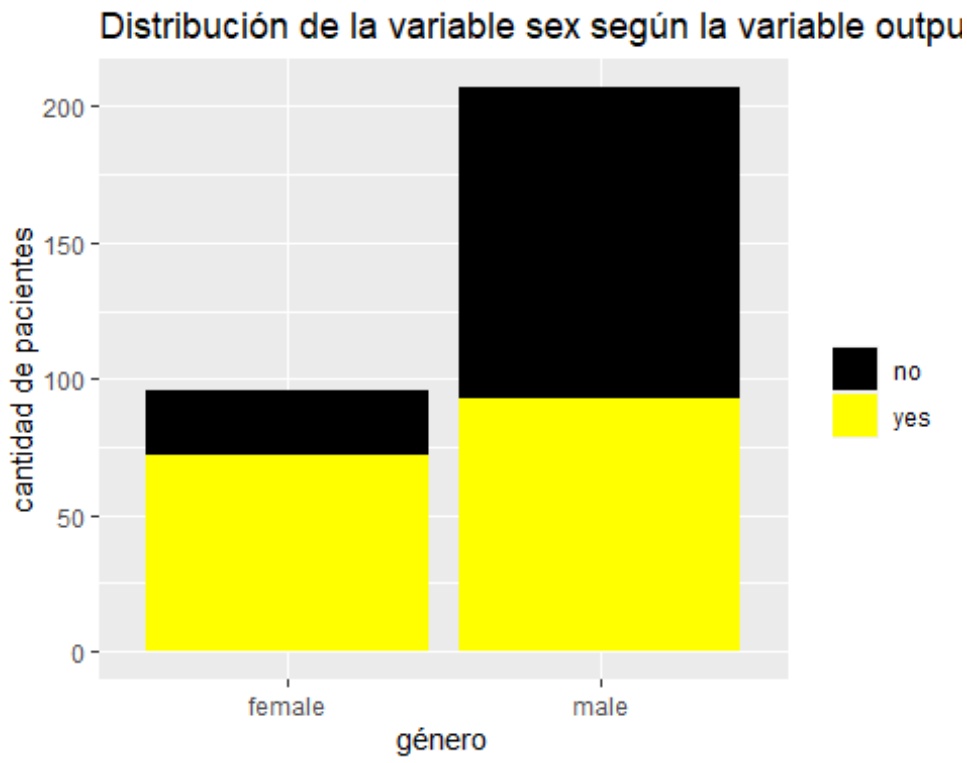



Del gráfico anterior podemos observar que hay más pacientes masculinos que no tienen problema de infarto que las pacientes femeninas.

A contianuación veremos el porcentaje de mujeres que tienen problemas de corazón y el porcentaje de hombres:

```
sexbyOutput<-ggplot(dat, aes(sex, fill=output))+geom_bar() +labs(x="género",
y="cantidad de pacientes")+ guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","yellow"))+ggtitle("Distribución de la
variable sex según la variable output")

grid.arrange(sexbyOutput, ncol=1)
```




Se observa que la gran parte de las mujeres tienen posibilidad alta de infarto mientras el porcentaje en los hombres no llega ni el 50%.

Del caso anterior podemos llegar a la conclusión de que factorizar las variables binarias nos puede facilitar el análisis del dataset, así que a continuación modificaremos las otras variables binarias:

```
dat["fbs"][dat["fbs"] == 0] <- 'false'
dat["fbs"][dat["fbs"] == 1] <- 'true'
dat$fbs <- as.factor(dat$fbs)

dat["exng"][dat["exng"] == 0] <- 'no'
dat["exng"][dat["exng"] == 1] <- 'yes'
dat$exng <- as.factor(dat$exng)
```

## 3. Limpieza de los datos.

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

Buscar si hay valores nulos o pérdidos:

```
colSums(is.na(dat))

##      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
```

```
colSums(dat=="")
```

```
##      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
```

Se observa que no hay valores ceros ni elementos vacíos, el dataset lo podemos considerar bueno para el análisis.

Ahora ya tenemos el dataset preparado para el análisis.

## 3.2. Identifica y gestiona los valores extremos.

Los valores extremos o outliers son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos utilizar la función boxplots.stats() de R recorriendo a cada una de las variables para encontrar los outliers:

```
boxplot.stats(dat$age)$out
```

```
## integer(0)
```

```
boxplot.stats(dat$sex)$out
```

```
## Warning in Ops.factor(x[floor(d)], x[ceiling(d)]): '+' not meaningful for
## factors
```

```
## factor(0)
## Levels: female male
```

```
boxplot.stats(dat$cp)$out
```

```
## integer(0)
```

```
boxplot.stats(dat$trtbps)$out
```

```
## [1] 172 178 180 180 200 174 192 178 180
```

Estos valores de la presión arterial son reales, aunque han superado del rango normal, se consideran por lo tanto Hipertensión arterial.

```
boxplot.stats(dat$chol)$out
```

```
## [1] 417 564 394 407 409
```

Estos valores de colesterol también existen, pueden corresponder a los pacientes que tienen la hipercolesterolemia.

```
boxplot.stats(dat$fbs)$out
```

```
## Warning in Ops.factor(x[floor(d)], x[ceiling(d)]): '+' not meaningful for
## factors

## factor(0)
## Levels: false true
```

Es una variable binaria, así que es normal tener 1 y 0.

```
boxplot.stats(dat$restecg)$out

## integer(0)

boxplot.stats(dat$exng)$out

## Warning in Ops.factor(x[floor(d)], x[ceiling(d)]): '+' not meaningful for
## factors

## factor(0)
## Levels: no yes

boxplot.stats(dat$slp)$out

## integer(0)

boxplot.stats(dat$caa)$out

##  [1] 3 4 3 3 4 4 4 3 3 3 3 3 3 3 3 3 3 3 3 4 3 3 3 3
```

caa es el número de vasos sanguineos principales, estos valores son posibles para esa variable.

```
boxplot.stats(dat$thall)$out

## [1] 0 0
```

0 significa que el paciente no tiene talasemia, son valores normales.

```
boxplot.stats(dat$thalachh)$out

## [1] 71
```

Es un valor poco habitual para frecuencia cardíaca máxima alcanzada, pero teniendo en cuenta que los pacientes del dataset son los que tienen problemas de corazón o posibles enfermedades de corazón, este valor puede ser real de un paciente.

```
boxplot.stats(dat$oldpeak)$out

## [1] 4.2 6.2 5.6 4.2 4.4
```

El valor promedio de la Depresión ST inducida por el ejercicio en relación al descanso es de 1.04, el valor mínimo es de 0 y el máximo de 6.20. Como están dentro del rango, no habrán problema.

Hemos encontrado los outliers para cada variable, pero después de haberlos analizado bien, son los valores que pueden tener los pacientes reales, por lo tanto, no hacemos nada con estos valores extremos para poder tener un resultado correcto.

# 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?)

Como el objetivo es para predecir la probabilidad del infarto de los pacientes, podemos agrupar los pacientes que tienen menos probabilidad de sufrir un infarto y los pacientes que tienen más probabilidad de sufrir el infarto para poder compararlos y conseguir la diferencia entre ellos.

Por otro lado, podemos agrupar los pacientes por el género (masculinos y femeninos) y también por edad (inferiores de 50 años y superiores de 50 años). De esta manera podremos obtener conclusiones sobre la influencia de la edad y el género sobre las enfermedades de corazón.

```
dat.attack <- dat[dat$output == 1,]
dat.healthy <- dat[dat$output == 0,]

dat.male <- dat[dat$sex == 1,]
dat.female <- dat[dat$sex == 0,]

dat.old <- dat[dat$age > 50,]
dat.young <- dat[dat$age < 51,]
```

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

Una forma de explorar la normalidad de un conjunto de datos es por medio de las pruebas de normalidad. Las hipótesis para este tipo de pruebas son:

H0: la muestra proviene de una población normal.

H1: la muestra NO proviene de una población normal.

Para ellos, podemos utilizar la Prueba Shapiro-Wilk con la función shapiro.test con un nivel de significancia de 0.05, es decir, Alfa=0.05.

Criterio de Decisión

- Si P < Alfa, Se rechaza H0
- Si p >= Alfa, No se rechaza H0

Aplicamos solo a las variables cuantitativas:

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

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$age
## W = 0.98637, p-value = 0.005798
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$trtbps
## W = 0.96592, p-value = 1.458e-06
```

```
shapiro.test(dat$chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$chol
## W = 0.94688, p-value = 5.365e-09
```

```
shapiro.test(dat$thalachh)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$thalachh
## W = 0.97632, p-value = 6.621e-05
```

```
shapiro.test(dat$oldpeak)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$oldpeak
## W = 0.84418, p-value < 2.2e-16
```

```
shapiro.test(dat$caa)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dat$caa
## W = 0.72812, p-value < 2.2e-16
```

Se observa que los valores de p son todos menores que alfa, así que se rechazan la hipótesis nula, por lo tanto, las variables cuantitativas no siguen una distribución normal.

A continuación, estudiaremos la homogeneidad de varianzas mediante la aplicación de un test de Fligner-Killeen. En este caso, estudiaremos la homogeneidad de varianzas de edad a

los grupos de pacientes con posibilidad alta de infarto y posiblidad baja de infarto. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```
fligner.test(age ~ output, data = dat)
```

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

Se observa que el valor p es menor que alfa (0,05), se rechaza la hipótesis nula, por lo tanto, las varianzas de ambas muestras no son homogéneas.

Ahora investigamos la homogeneidad entre la variable chol y la variable output:

```
fligner.test(chol ~ output, data = dat)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chol by output
## Fligner-Killeen:med chi-squared = 0.63573, df = 1, p-value = 0.4253
```

Se observa que el valor p es mayor que alfa, por lo tanto, se acepta la hipótesis nula 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.

### 4.3.1. Modelo de Regresión logística

Tal y como se planteó en los objetivos de la actividad, resultará de mucho interés poder realizar predicciones sobre la probabilidad alta o baja del infarto de un paciente teniendo los valores de los factores. De esta forma, se calculará un modelo de regresión logística utilizando regresores tanto cuantitativos como cualitativos con el que poder realizar las predicciones de la probabilidad alta (1) o baja (0).

```
log_mod <-
glm(output~age+sex+cp+trtbps+chol+fbs+restecg+thalachh+exng+oldpeak+slp+caa+t
hall, family = binomial(), data = dat)
summary(log_mod)
```

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

```
##
## 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
## sexmale     -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
## fbstrue      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 *
## exngyes     -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
```

Se observa que las 7 de las 13 variables incluidas en el modelo son significativas puesto que sus valores de p son menores que 0,05, las cuales son sex (male), cp, thalachh, exng (yes), oldpeak, caa, thall.

Donde las variables cp, thalachh tienen una correlación positiva con la variable output, indicando que cuanto mayores sean los valores de estas variables (dolor de pecho no anginoso o aasintomático, frecuencia cardíaca máxima alcanzada mayor), mayor posibilidad de infarto tiene ese paciente. Por el contrario, la variable sex, exng, oldpeak, caa, thall tienen una correlación negativa: si el género es hombre, sí tiene angina inducida por el ejercicio, depresión del ST inducida mayor, número de vasos sanguíneos principales mayor, talasemia normal o defecto reversible, su posibilidad de infarto es menor.

La calidad del modelo podemos medir comparando null deviance con residual deviance, de la siguiente forma:

(null.deviance - deviance)/null.deviance

```
indice <- (417.64 - 211.44) / 417.64
indice

## [1] 0.4937267
```

El resultado indica si el modelo tiene un ajuste perfecto o no, cuanto mayor es el valor (más cerca a 1) mejor es el ajuste. En nuestro caso, se obtiene: 0.49, que no tiene un ajuste perfecto pero tampoco es un ajuste malo.

### 4.3.2. Contraste de hipótesis

Primero formulamos la pregunta de investigación:

¿La proporción de los pacientes que tienen posibilidad alta de infarto es mayor entre las mujeres que entre los hombres?

Escribid la hipótesis nula y la hipótesis alternativa.

- H0 : pF = pM
- H1 : pF > pM

Para obtern la respuesta de la pregunta podemos aplicar un test de hipótesis de dos muestras sobre la media utilizando la función prop.test.

```
female <- dat[which(dat$sex == "female"), "output"]
data_f <- as.array(female)
nf <- nrow(data_f)
male <- dat[which(dat$sex == "male"), "output"]
data_m <- as.array(male)
nm <- nrow(data_m)

nATinF <- nrow(data_f[which(data_f=="yes")])
pinF <- nATinF / nf
nATinM <- nrow(data_m[which(data_m=="yes")])
pinM <- nATinM / nm

successCount <- c(nATinF, nATinM)
n <- c(nf, nm)
res <- prop.test(successCount, n, alternative="greater", correct=FALSE)
res

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  successCount out of n
## 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 valor p (5.036e-07) es menor que alfa (0.05), Se concluye que nos encontramos en la zona de rechazo de la hipotesis nula, con lo cual, podemos afirmar que la proporción de los pacientes que tienen posibilidad alta de infarto es mayor entre las mujeres que entre los hombres.

Lo cual es muy lógico, puesto que podemos comprobar con las proporciones calculadas (0,750 en mujeres y 0,449 en hombre).

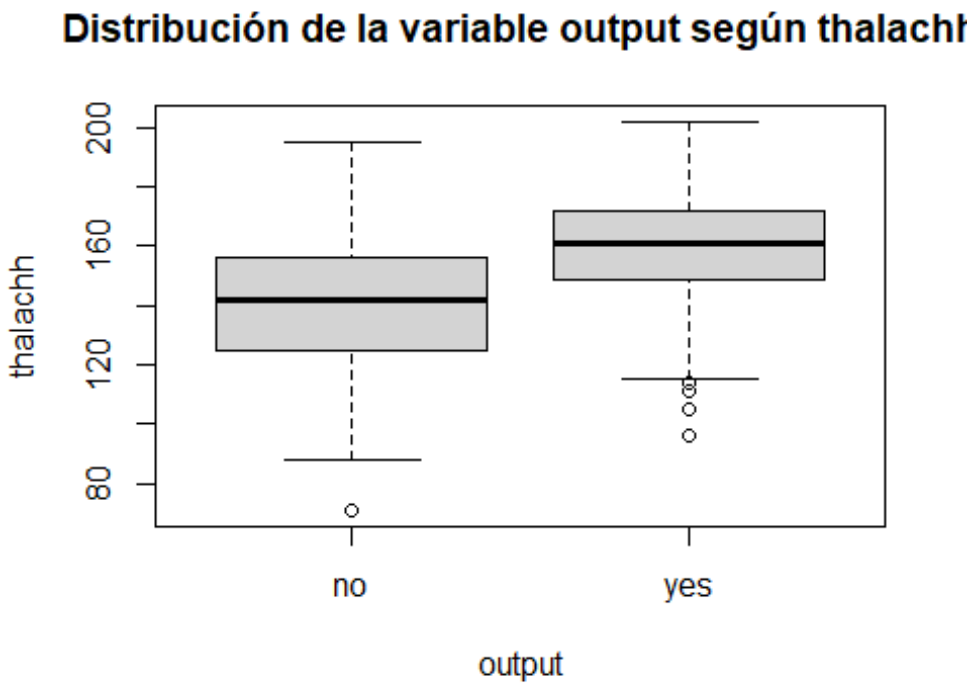
### 4.3.3. Análisis de la varianza (ANOVA)

Vamos a realizar un ANOVA para contrastar si existen diferencias en la variable thalach en función de la posibilidad alta o baja de infarto de los pacientes.

Visualización gráfica

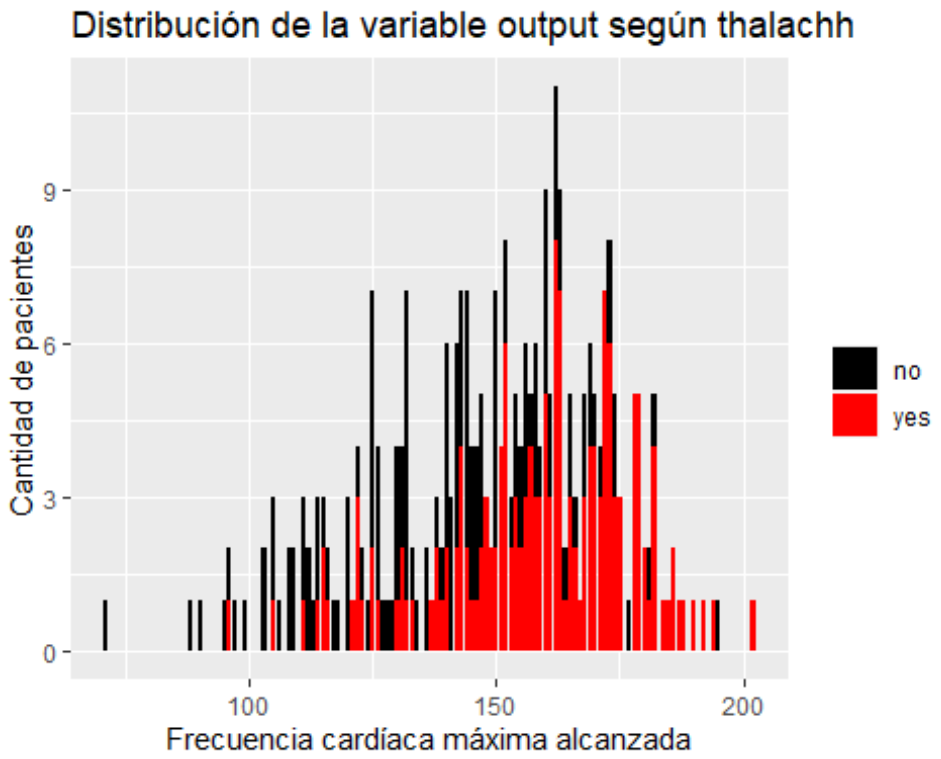
Mostrad gráficamente la distribución de output según los valores de thalach.

```
boxplot(dat$thalachh~dat$output, main="Distribución de la variable output
según thalachh",xlab="output",ylab="thalachh")
```




```
outputbyThalachh<-ggplot(dat, aes(thalachh, fill=output))+geom_bar()
+labs(x="Frecuencia cardíaca máxima alcanzada", y="Cantidad de pacientes")+
guides(fill=guide_legend(title=""))+
scale_fill_manual(values=c("black","red"))+ggtitle("Distribución de la
variable output según thalachh")
```

```
grid.arrange(outputbyThalachh, ncol=1)
```




Hipótesis nula y alternativa

Formulamos la pregunta:

¿La frecuencia cardíaca máxima alcanzada es igual entre los pacientes que tienen alta posibilidad de infarto y los que tienen baja posibilidad de infarto?

- H0: μYes = μNo
- H1: μYes ≠ μNo

Calculamos el análisis de varianza usando la función aov o lm:

Primero usamos la función aov:

```
mod_aov <- aov(thalachh~output, dat)
summary(mod_aov)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## output        1  28182   28182   65.12 1.7e-14 ***
## Residuals   301 130262     433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ahora usamos la función lm:

```
mod_lm <- lm(thalachh~output,data=dat)
anova(mod_lm)

## Analysis of Variance Table
##
## Response: thalachh
##            Df Sum Sq Mean Sq F value    Pr(>F)
## output      1  28182 28181.6   65.12 1.697e-14 ***
## Residuals 301 130262   432.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con las dos funciones hemos sacado los mismos resultados:

Valores del contraste: Sum Sq = 28182; Mean Sq = 28181.6; estadístico F = 65.12; valor de p = 1.7e-14.

El valor de p es menor que 0.05, podemos decir que el factor analizado es significativo y nos encontramos en la zona de rechazo de la hipótesis nula.

En conclusión, La frecuencia cardíaca máxima alcanzada es diferente entre los pacientes que tienen alta posibilidad de infarto y los que tienen baja posibilidad de infarto.

## 5. Resolución del problema

Hemos realizado tres tipos de pruebas estadísticas sobre un conjunto de datos que se correspondía con diferentes factores correspondientes a pacientes con el motivo de cumplir en la medida de lo posible con el objetivo que se planteaba al comienzo. Para cada una de ellas, hemos podido ver la relación entre las variables con la variable objetivo (output). Así como el modelo de regresión logística nos ha permitido conocer cuáles de estas variables ejercen una mayor influencia sobre la posibilidad de sufrir un infarto. El contraste de hipótesis y el modelo de ANOVA nos ha permitido conocer la relación entre una variable en concreto con la variable objetivo. Por lo tanto, los resultados nos han permitido responder al problema inicialmente planteado.

Previamente, se han sometido los datos a un preprocesamiento para factorizar las variables binarias para un mejor análisis y visualización posteriores.

El dataset con los datos finales analizados se lo guardamos en un archivo .csv nuevo.

```
write.csv(dat, "final_heart.csv")
```

## 6. Contribuciones:

Una única autora Fanfan Yang, contribución 100%.