UOC - Tipologia y ciclo de vida de los datos - PRA2

Limpieza y Preprocesado: Heart Attack Analysis & Prediction Dataset

Vanessa Moreno González, Manuel Ernesto Martínez Martín

4 de June 2023

Índice

1. Descripción del dataset

Este dataset trae dos ficheros `heart.csv` y `o2Saturation.csv` y **es importante porque proporciona información sobre factores relacionados con enfermedades cardíacas**, como edad, sexo, síntomas otros datos médicos. Ya que con el **se puede entender mejor la enfermedad y hacer un análisis para detectar cuando se puede estar en riesgo de ataque cardíaco**, sabiendo esto se pueden desarrollar modelos predictivos que tomen decisiones para ayudar a prevenir un ataque cardíaco.

El dataset es el propuesto en el enunciado de la práctica y se ha extraído de kaggle: **Heart Attack Analysis & Prediction Dataset**

Contenido del dataset

Las variables que tiene el dataset son: age, sex, cp, trtbps, chol, fbs, restecg, thalachh, exng, oldpeak, slp, caa, thall y output. Siendo `output` la variable objetivo. A continuación se detallan más en profundidad.

- **age**: Edad del paciente.
- **sex**: Género del paciente.
  - *0*: Femenino
  - *1*: Masculino
- **cp**: Tipo de dolor en el pecho.
  - *0*: Angina típica
  - *1*: Angina atípica
  - *2*: Dolor no anginal
  - *3*: Asintomático
- **trtbps**: Presión arterial en reposo (en mm Hg).
- **chol**: Colesterol en mg/dl medido mediante un sensor BMI.
- **fbs**: Nivel de azúcar en sangre en ayunas (> 120 mg/dl).
  - *1*: Verdadero
  - *0*: Falso
- **restecg**: Resultados electrocardiográficos en reposo.
  - *0*: Normal
  - *1*: Anormalidad con inversiones de onda ST-T y/o alteraciones del segmento ST > 0.05 mV
  - *2*: Hipertrofia ventricular izquierda
- **thalachh**: Ritmo cardíaco máximo alcanzado.
- **exng**: Angina inducida por ejercicio.
  - *1*: Sí
  - *0*: No
- **oldpeak**: Diferencia entre la depresión del segmento ST durante el ejercicio y durante el descanso en un electrocardiograma.
- **slp**: Pendiente del segmento ST durante el ejercicio en la prueba de esfuerzo.
  - *1*: Ascendente
  - *2*: Plana
  - *3*: Descendente
- **caa**: Número de vasos principales (0-3).
- **thall**: Talasemia, trastorno hereditario de la sangre caracterizado por un menor nivel de hemoglobina.
  - *0*: Ausencia

  - *1*: Talasemia normal
  - *2*: Talasemia fija defectuosa
  - *3*: Talasemia Reversible defectuosa

- **output**: Variable objetivo.
  - *0*: Menor probabilidad de ataque al corazón
  - *1*: Mayor probabilidad de ataque al corazón

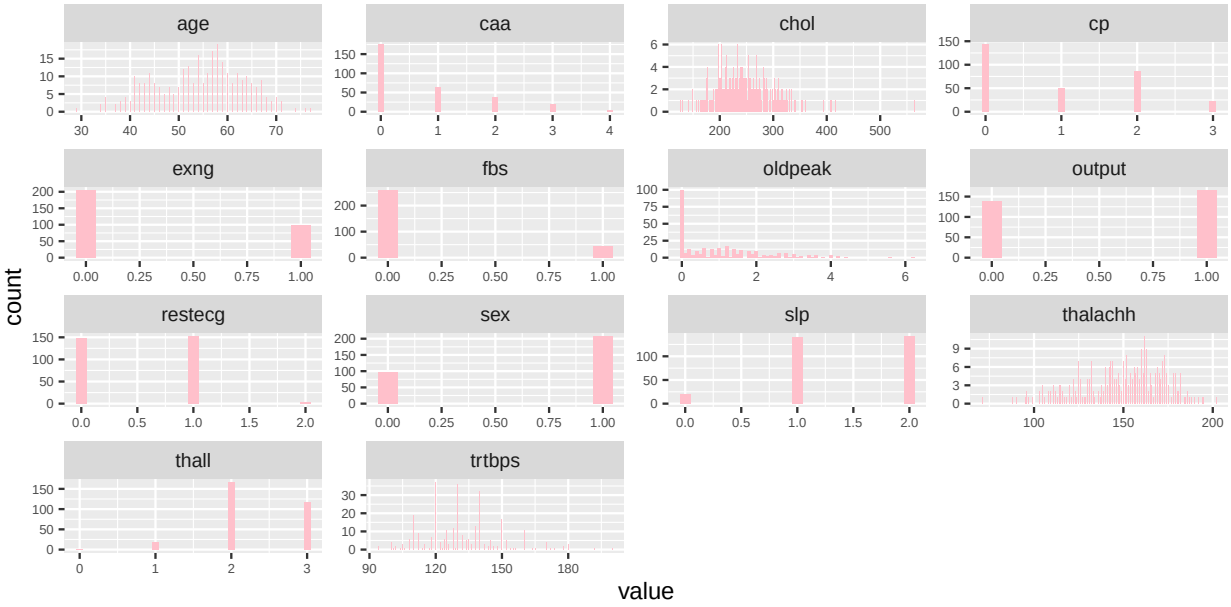
Análisis inicial

Verificamos la estructura del juego de datos principal y el tipo de datos con los que R ha interpretado cada variable, y si, corresponde a la descripción de las variables del fichero original (podemos usar `str()` o `glimpse()`):

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

Observamos que todas las variables se han cargado como númerica discreta a excepción de **oldpeak** que se ha cargado como numérica continua.

A continuación mostraremos una visión en general con gráficas de los valores de las variables numéricas para determinar la cantidad de valores que pueden tener, esto nos ayuda también a confirmar cuales son numéricas continuas o cuales son numéricas discretas



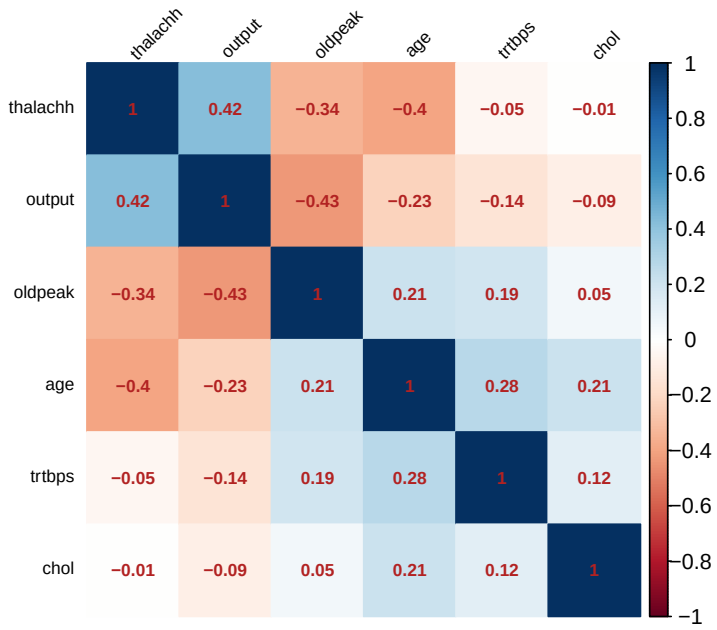
Observamos como **sex, caa, cp, fbs, restecg, exng, slp y thall** contienen un número limitado de valores únicos, por lo que, probablemente, estén representando variables categóricas. Comprobando, según la descripción oficial del dataset del punto anterior, que es cada variable y si nuestro análisis inicial es correcto, observamos que nuestra suposición es correcta, y que son variables categóricas. Por lo tanto, las convertiremos usando la función `mutate()`:

2. Integración y selección de variables

Después de análisis inicial, vemos que hay realmente dos ficheros que no estan alineados en el número de registros y que tampoco tenemos forma de unirlos, **`heart.csv`** tiene **14 variables** y **303 registros** mientras que **`o2Saturation.csv`** con **1 variable** y **3585 registros**.

Aunque el nivel de saturación de oxigeno pueda ser importante para los ataques cardíacos, no hay manera de juntar los dos conjuntos de datos en uno solo debido a que no hay un identificador de paciente, por lo que solo usaremos **`heart.csv`**.

Para la selección de los datos, comprobaremos la correlación entre ellas. En el caso de las variables numéricas, realizaremos la correlación de Pearson:



Tanto una correlación positiva como una muy negativa son interesantes para la selección de variables. Centrándonos en la fila de la variable objetivo `output` se tienen los siguientes valores:

```
##      age   trtbps     chol thalachh  oldpeak   output
##    -0.23    -0.14    -0.09     0.42    -0.43     1.00
```

Se puede tomar como referencia **0.1** como umbral para comprobar las variables que no son necesarias para el estudio, siempre en valor absoluto. En este caso para el coeficiente de correlación de pearson se tienen `age`, `trtbps`, `thalachh` y `oldpeak` como variables aptas y `chol` como poco importante.

Para las variables categóricas numéricas seria más apropiado hacer un test de `Fisher` o un `Chi-squared`.

Se va a proceder a hacer uso del test de `Fisher` con `fisher.test()`

$p = \frac{\binom{a+b}{a} \cdot \binom{c+d}{c}}{\binom{n}{a+c}}$

Visualizamos los resultados obtenidos del test de Fisher:

```
## $sex
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 1.042e-06
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1519598 0.4783553
## sample estimates:
## odds ratio
##  0.2731136
##
##
## $caa
##
```

```
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value < 2.2e-16
## alternative hypothesis: two.sided
##
##
## $cp
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value < 2.2e-16
## alternative hypothesis: two.sided
##
##
## $fbs
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 0.6308
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4308961 1.6975867
## sample estimates:
## odds ratio
##  0.8544825
##
##
## $restecg
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 0.003629
## alternative hypothesis: two.sided
##
##
## $exng
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 1.76e-14
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.07259027 0.23708719
## sample estimates:
## odds ratio
##   0.133146
##
##
## $slp
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value = 1.165e-11
## alternative hypothesis: two.sided
##
##
## $thall
##
##  Fisher's Exact Test for Count Data
##
## data:  heartAttack[[var]] and heartAttack$output
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Se entiende entonces que las variables que tienen un *p-valor* por debajo de un nivel de significancia de **0.05** son consideradas buenas para ser escogidas para el análisis, es decir estas variables tienen un buen nivel estadístico de significancia y aportan información a los posibles modelos en las que se incluyan. De las variables categóricas seleccionadas todas menos `fbs` tienen un p-valor por debajo de `0.05`.

Puesto que `fbs` no es una variable significativa se va a evitar su uso.

3. Limpieza de los datos

Volvemos a comprobar la estructura de los datos con `str()`, para verificar que los cambios realizados anteriormente se han ejecutado correctamente. Además mostraremos el resumen general de cada una de las variables con sus valores máximos, mínimos media, mean y cuartiles utilizando la función `summary()`. Es aquí donde en los casos numéricos se pueden ver si hay valores imposibles de cumplir tanto en máximos como en mínimos.

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 2 ...
##  $ cp      : Factor w/ 4 levels "0","1","2","3": 4 3 2 2 1 1 2 2 3 3 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
##  $ restecg : Factor w/ 3 levels "0","1","2": 1 2 1 2 2 2 1 2 2 2 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : Factor w/ 3 levels "0","1","2": 1 1 3 3 3 2 2 3 3 3 ...
##  $ caa     : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ thall   : Factor w/ 4 levels "0","1","2","3": 2 3 3 3 3 2 3 4 4 3 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
##       age        sex     cp          trtbps           chol       fbs
##  Min.   :29.00   0: 96   0:143   Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:47.50   1:207   1: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :55.00           2: 87   Median :130.0   Median :240.0
##  Mean   :54.37           3: 23   Mean   :131.6   Mean   :246.3
```

```
##  3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:274.5
##  Max.   :77.00                   Max.   :200.0   Max.   :564.0
##  restecg    thalachh     exng       oldpeak     slp     caa     thall
##  0:147   Min.   : 71.0   0:204   Min.   :0.00   0: 21   0:175   0:  2
##  1:152   1st Qu.:133.5   1: 99   1st Qu.:0.00   1:140   1: 65   1: 18
##  2:  4   Median :153.0           Median :0.80   2:142   2: 38   2:166
##          Mean   :149.6           Mean   :1.04           3: 20   3:117
##          3rd Qu.:166.0           3rd Qu.:1.60           4:  5
##          Max.   :202.0           Max.   :6.20
##      output
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :1.0000
##  Mean   :0.5446
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

Como se puede observar las variables categóricas ya están en tipo factor.

De la variable `caa` se tenían identificados valores de 0 a 3, **pero el valor máximos es 4**.

Nota: este valor de 4 en `caa` será eliminado en el apartado de los valores extremos.

3.1. ¿Los datos contienen ceros o elementos vacíos?

Cuando en un dataset se tienen datos nulos, hay una serie de estrategias a seguir para solucionar esto y que el juego de datos se pueda usar:

- **Eliminación de los registros**, esto a veces no es adecuado porque puede perderse mucha información que hay en otras variables que pueden ser más importantes.
- **Imputación de un valor** que puede ser: utilizar la media, la mediana, la moda, interpolación, utilización de los vecinos cercanos, u otros métodos.

Búsqueda de ceros

Tenemos algunas variables categóricas en formato numérico en nuestro conjunto de datos. Estas variables no se pueden considerar en la búsqueda de ceros, ya que el valor `0` es una de las posibles categorías para cada una de ellas. Las variables categóricas en formato numérico que ahora son factor son `sex`, `cp`, `fbs`, `restecg`, `exng`, `slp`, `thall` y la target `output`. De las cuales son dicotomicas `sex`, `fbs`, `exng` y `output`. Además existe la variabe `caa` con tres posibles valores que indican una cantidad que puede ser 0.

También en el resumen mostrado anterior se podía ver a simple vista si alguna variable tenía 0 si este fuera su valor mínimo.

Para buscar los valores con ceros podemos usar `colSums()` y comprobando con un `=` como a continuación

```
##      age   trtbps     chol thalachh  oldpeak
##        0        0        0        0       99
```

- **age**: Hay 0 pacientes con 0 años.
- **trtbps**: Hay 0 pacientes con 0 o sin presión arterial en reposo.
- **chol**: Hay 0 pacientes con 0 o sin medición de colesterol.
- **thalachh**: Hay 0 pacientes con 0 o sin ritmo cardíaco máximo alcanzado.
- **oldpeak**: Hay 99 pacientes con 0 o sin informar de la diferencia en segmento ST con electrocardiograma.

Búsqueda de NAs

Para buscar los valores nulos podemos usar de nuevo `colSums()` pero ahora con `is.na()`

```
##      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
```

Como se puede observar **no hay valores `NA`** en este dataset, otra comprobación sería buscar valores en blanco, pero esto se haría si hubiera variables categóricas que fueran cadenas, en este caso no es necesario ya que no hay ningún valor como texto.

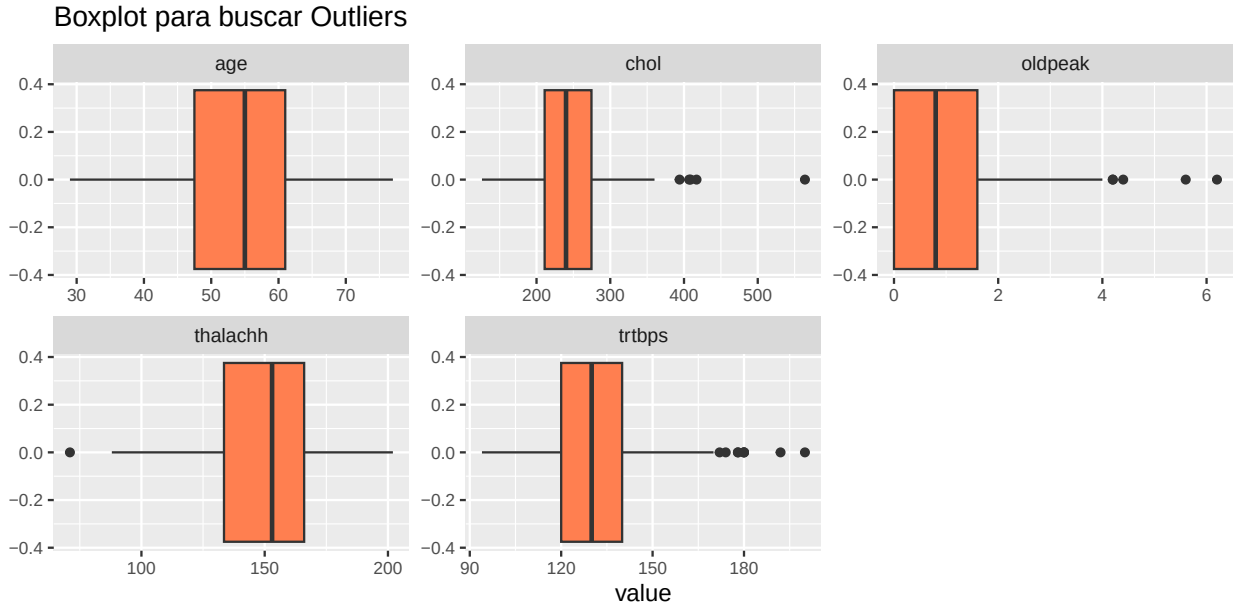
3.2. Identifica y gestiona los valores extremos

Primero en las variables numéricas vamos a comprobar cuantos valores atípicos hay de cada una

```
##      age   trtbps     chol thalachh  oldpeak
##        0        9        5        1        5
```

Como se puede observar en las variables numéricas, `age` no tiene valores atípicos.

Ahora se visualizarán los valores atípicos de las variables numéricas.



También queremos ver si se va a perder mucha información a la hora de borrar registros por ello veremos cuantos registros tenemos a tes y después de la eliminación

```
## Valores atípicos de 'chol': 394 407 409 417 564 con un total de 5 registros
```

```
## Valores atípicos de 'oldpeak': 4.2 4.4 5.6 6.2 con un total de 5 registros
```

```
## Valores atípicos de 'thalachh': 71 con un total de 1 registros
```

```
## Valores atípicos de 'trtbps': 172 174 178 180 192 200 con un total de 9 registros
```

Se opta por eliminar los registros con valores atípicos

También aunque `caa` no se ha tenido en cuenta para comprobar los valores atípicos porque es una variable categórica, lo cierto es que según la información del dataset solo tiene 4 categorías representadas por los valores 0, 1, 2 o 3. pero existen 5 registros donde se tiene una 5^a categoría y debería ser eliminada. Como es de tipo factor se debe de usar `droplevels()` para que se elimine el nivel de dicha categoría.

```
## Cantidad de registros antes 303 y después de eliminar los valores atípicos 279
```

Realmente de 303 registros a 279 no hay mucha diferencia así que se mantiene la eliminación de outliers.

4. Análisis de los datos

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

Deseamos conocer la relación que existe entre las siguientes variables:

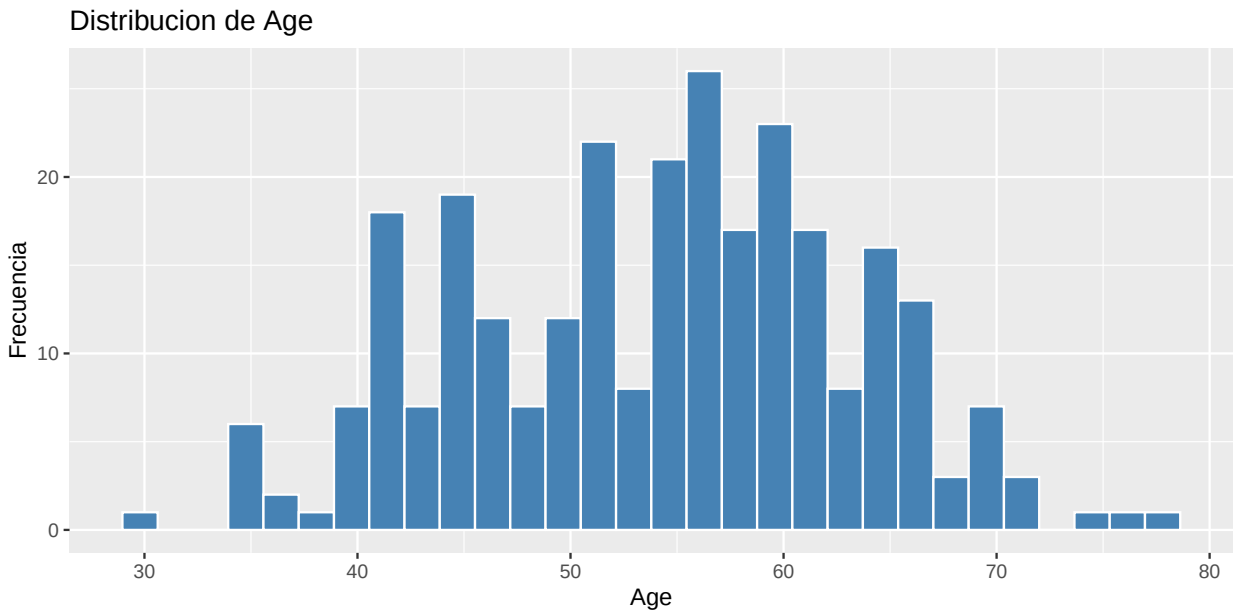
- **sex**, **cp**: Queremos conocer si existen diferencias significativas entre el tipo de dolor en el pecho que experimentan las observaciones con infarto en función del sexo.
- **age**, **trtbps** : Queremos conocer si la media de la variable numérica **trtbps** es la misma para los grupos de datos **age** tras realizar una discretización de esta variable.
- **output vs variables**: Queremos aproximar la relación de dependencia que existe entre la probabilidad de sufrir un infarto y las variables del dataset.

4.2. Comprobación de la normalidad y homogeneidad de la varianza

sex, cp y output: No es necesario comprobar la normalidad y homogeneidad de la varianza, ya que aplicaremos el test chi-cuadrado, que se trata de un test no paramétrico.

age: Seleccionaremos el test a aplicar en función de la normalidad y la homogeneidad de la varianza de los grupos a comparar.

El primer paso es discretizar la variable age. Primero vamos a mostrar un histograma de la distribución de las edades



Y a continuación, discretizamos la variable en tres grupos: `Jovenes`, `Media` y `Vejez`, de 0 a 45, de 45 a 60 y 60 a 80

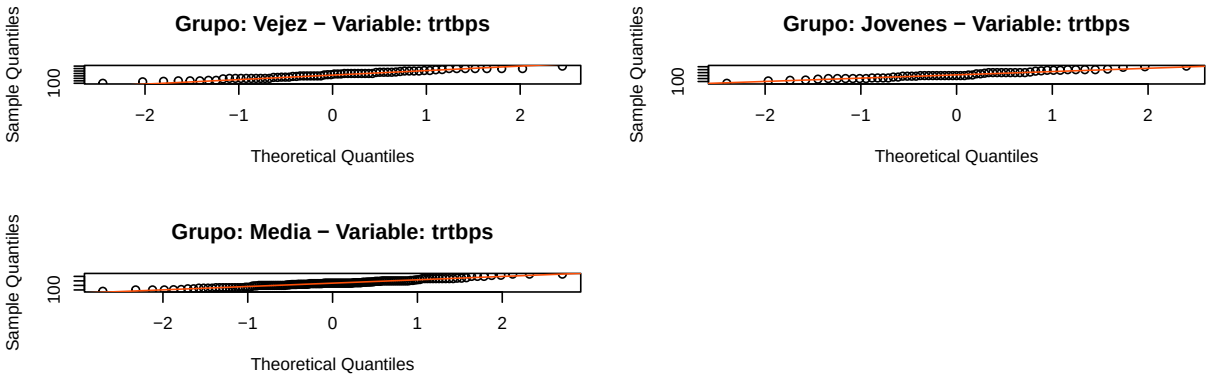
Visualizamos el mínimo y máximo valor para age de cada grupo.

```
## Jovenes - Mínimo: 29 Máximo: 45
```

```
## Medio - Mínimo: 46 Máximo: 60
```

```
## Vejez - Mínimo: 61 Máximo: 77
```

Ahora comprobamos si la variable `trtbps`, presenta una distribución normal de manera visual, a través del gráfico Q-Q:



De manera visual, parece que la variable `trtbps` se aleja de la recta normal en la parte derecha el gráfico en el grupo edad Media.

Ahora realizaremos una evaluación más cuantitativa de la distribución, mediante el test Shapiro-Wilk con `shapiro.test()`:

```
## Grupo: Vejez
## El p-value de trtbps es: 0.2293937
## La variable 'trtbps' en el grupo Vejez sigue una distribución normal.
##
## Grupo: Jovenes
## El p-value de trtbps es: 0.2786165
## La variable 'trtbps' en el grupo Jovenes sigue una distribución normal.
##
## Grupo: Media
## El p-value de trtbps es: 0.03323032
## La variable 'trtbps' en el grupo Media no sigue una distribución normal.
```

Observamos como se cumple lo que hemos visualizado en los gráficos Q-Q y para el grupo edad media, `trtbps` no presenta una distribución normal. Por lo que, descartamos el test de ANOVA y aplicaremos Kruskal-Wallis que no asume una distribución normal en los datos.

No es necesario comprobar la homogeneidad de los datos ya que, al no cumplir el criterio de normalidad, aplicaremos un test no paramétrico.

output vs variables: Aproximaremos la relación de dependencia entre la variable dependiente output y el resto de variables. En este caso, no es necesario que las variables presenten una distribución normal, ya que emplearemos la regresión logística, y estos supuestos no son requisitos para el modelo.

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

CHI-CUADRADO

Para analizar las diferencias entre el tipo de dolor en el pecho y el sexo en las observaciones con infarto realizaremos la prueba de chi-cuadrado.

El primer paso es seleccionar únicamente las observaciones con output 1.

Calculamos, la frecuencia, en una tabla, con el tipo de dolor en el pecho y el sexo.

```
##
##      0  1
##   0 16 21
##   1 16 24
##   2 32 31
##   3  4 11
```

Realizamos la prueba chi-cuadrado:

```
##
##  Pearson's Chi-squared test
##
## data:  tabla
## X-squared = 3.2784, df = 3, p-value = 0.3507
```

El resultado de p-value superior a 0.05 indica que **no se encuentran diferencias significativas para el tipo de dolor en el pecho y el sexo dentro de la población que sufre un infarto**.

KRUSKAL-WALLIS

Para comparar la variable **trtbps** entre los grupos de edad **Jóvenes, Media y Vejez**, y como sabemos que no se cumple en supuesto de normalidad de los datos en cada grupo, aplicaremos Kruskal_Wallis.

Nuestras hipótesis son:

- Hipótesis nula (HO): No hay diferencias significativas de la media del valor trtbps entre los diferentes grupos de edad.
- Hipótesis alternativa (H1): Hay diferencias significativas de la media del valor trtbps entre los diferentes grupos de edad.

Realizamos el test de Kruskal-Wallis para la presión arterial y los diferentes grupos de edad:

```
##
##  Kruskal-Wallis rank sum test
##
## data:  trtbps by age_discretized
## Kruskal-Wallis chi-squared = 18.318, df = 2, p-value = 0.0001053
```

Los valores de p obtenidos, son inferiores a 0.05, por lo que podemos rechazar la hipótesis nula y concluir que **el valor de presión sanguínea varia en función del grupo de edad** de las observaciones.

REGRESIÓN LOGISTICA

Por último, aproximaremos la relación de dependencia entre la variable dependiente **output** y **el resto de variables** mediante una regresión logística.

Dividimos los datos entre train y test con las siguientes proporciones: 70% entrenamiento y 30% test.

Ajustamos el modelo de regresión logística y visualizamos el resultado del modelo:

```
##
## Call:
## glm(formula = output ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8060  -0.2088   0.0537   0.3094   2.9412
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -6.631e+00  5.700e+00  -1.163 0.244732
## age                  8.915e-02  8.203e-02   1.087 0.277144
## sex1                -9.851e-01  7.650e-01  -1.288 0.197872
## cp1                  6.782e-01  7.772e-01   0.873 0.382906
## cp2                  2.537e+00  7.872e-01   3.223 0.001270 **
## cp3                  1.355e+00  1.034e+00   1.310 0.190256
## trtbps              -1.588e-02  1.888e-02  -0.841 0.400239
## chol                -5.868e-04  6.912e-03  -0.085 0.932349
## fbs1                 8.792e-01  8.627e-01   1.019 0.308148
## restecg1             5.046e-01  5.552e-01   0.909 0.363431
## restecg2             9.951e+00  1.455e+03   0.007 0.994544
```

```
## thalachh              2.938e-02  1.904e-02   1.543 0.122816
## exng1                -4.653e-01  6.793e-01  -0.685 0.493354
## oldpeak              -1.016e+00  3.676e-01  -2.763 0.005733 **
## slp1                 -4.927e-02  1.129e+00  -0.044 0.965174
## slp2                  9.878e-01  1.172e+00   0.843 0.399228
## caa1                 -3.578e+00  7.979e-01  -4.484 7.32e-06 ***
## caa2                 -4.331e+00  1.177e+00  -3.679 0.000234 ***
## caa3                 -3.604e+00  1.379e+00  -2.612 0.008989 **
## thall1                2.663e+00  2.286e+00   1.165 0.243907
## thall2                3.175e+00  2.212e+00   1.435 0.151205
## thall3                6.361e-01  2.167e+00   0.294 0.769096
## age_discretizedMedia -1.071e+00  1.239e+00  -0.864 0.387493
## age_discretizedVejez -1.443e-01  2.078e+00  -0.069 0.944646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.843  on 194  degrees of freedom
## Residual deviance:  98.763  on 171  degrees of freedom
## AIC: 146.76
##
## Number of Fisher Scoring iterations: 14
```

Observamos que las variables que son estadísticamente significativas para output son **cp, oldpeak y caa**. Por lo tanto, volvemos a realizar un modelo con solo esas variables.

```
##
## Call:
## glm(formula = output ~ cp + oldpeak + caa, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5423  -0.5089   0.2074   0.5691   2.1412
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.2937     0.4012   3.225 0.001261 **
## cp1           1.7425     0.6193   2.814 0.004896 **
## cp2           2.5354     0.5728   4.427 9.58e-06 ***
## cp3           1.8978     0.7871   2.411 0.015907 *
## oldpeak      -1.1598     0.2476  -4.684 2.82e-06 ***
## caa1         -2.6510     0.5402  -4.908 9.22e-07 ***
## caa2         -2.1125     0.6731  -3.138 0.001699 **
## caa3         -3.3442     0.9908  -3.375 0.000737 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 268.84  on 194  degrees of freedom
## Residual deviance: 145.46  on 187  degrees of freedom
```

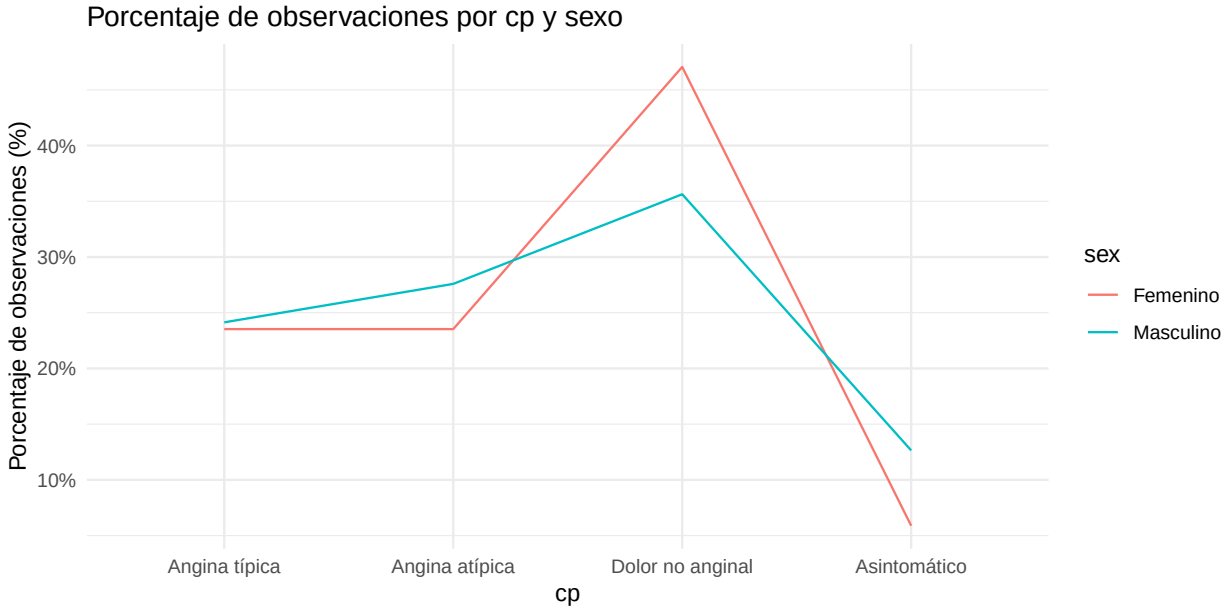
```
## AIC: 161.46
##
## Number of Fisher Scoring iterations: 5
```

La representación de los resultados de la regresión logística se mostrarán en el apartado 5.

5. Representación de los resultados

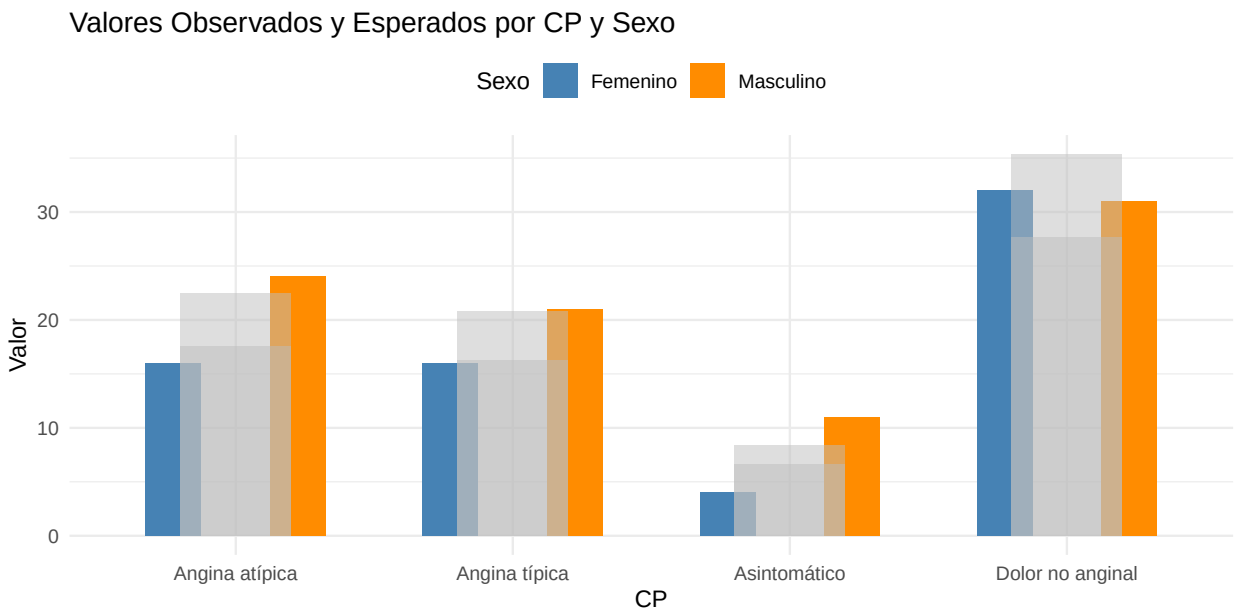
SEX-CP

El primer paso es representar gráficamente la cantidad de observaciones porcentualmente que han padecido infarto, agrupadas por sexo y por tipo de síntoma padecido.



Observamos como, para ambos sexos se experimentan, en orden de mayor a menor, Dolor no anginal, Angina atípica, Angina típica y Asintomático. Los porcentajes entre sexos varían, hayándose la mayor diferencia entre sexos en Dolor no Anginal, donde las mujeres lo experimentan en mas 10% por encima de los hombres.

El test Chi-cuadrado ha evaluado si existe una diferencia significativa entre estas dos variables categóricas, y para ello evalúa la diferencia entre las frecuencias observada y las frecuencias esperadas, en el supuesto nulo de independencia. Ahora vamos a gráficas los valores observados y los esperados.

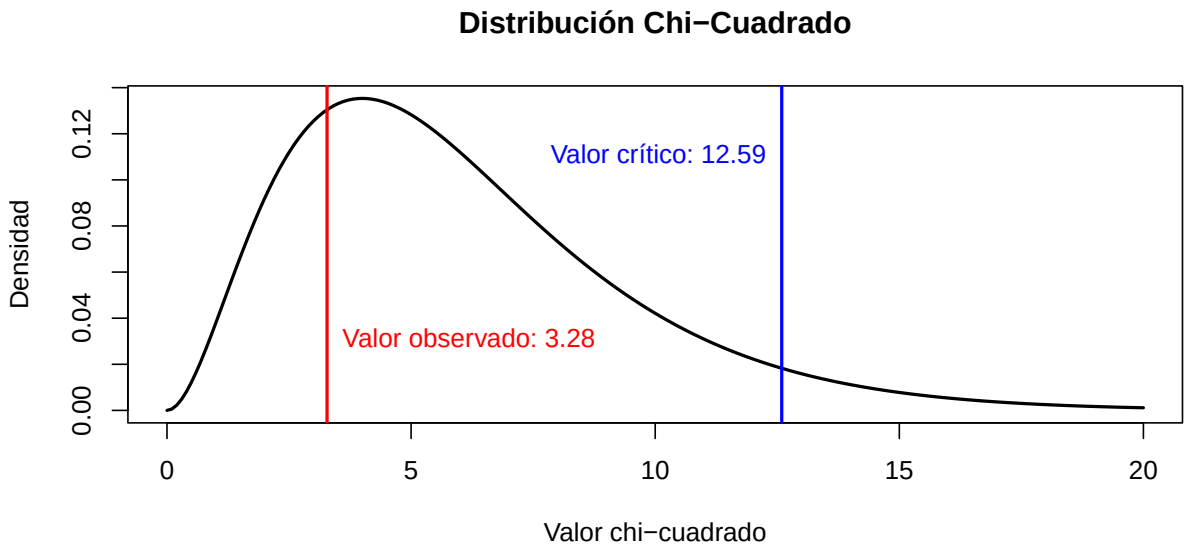


Visualizamos la tabla de valores:

Cuadro 1: Tabla de valores

cp	sexo	observados	esperados
Angina típica	Femenino	16	16.232258
Angina atípica	Femenino	16	17.548387
Dolor no anginal	Femenino	32	27.638710
Asintomático	Femenino	4	6.580645
Angina típica	Masculino	21	20.767742
Angina atípica	Masculino	24	22.451613
Dolor no anginal	Masculino	31	35.361290
Asintomático	Masculino	11	8.419355

Por último, graficamos la distribución de los datos Chi-cuadrado, dibujando el umbral crítico y nuestro resultado del test Chi-cuadrado.

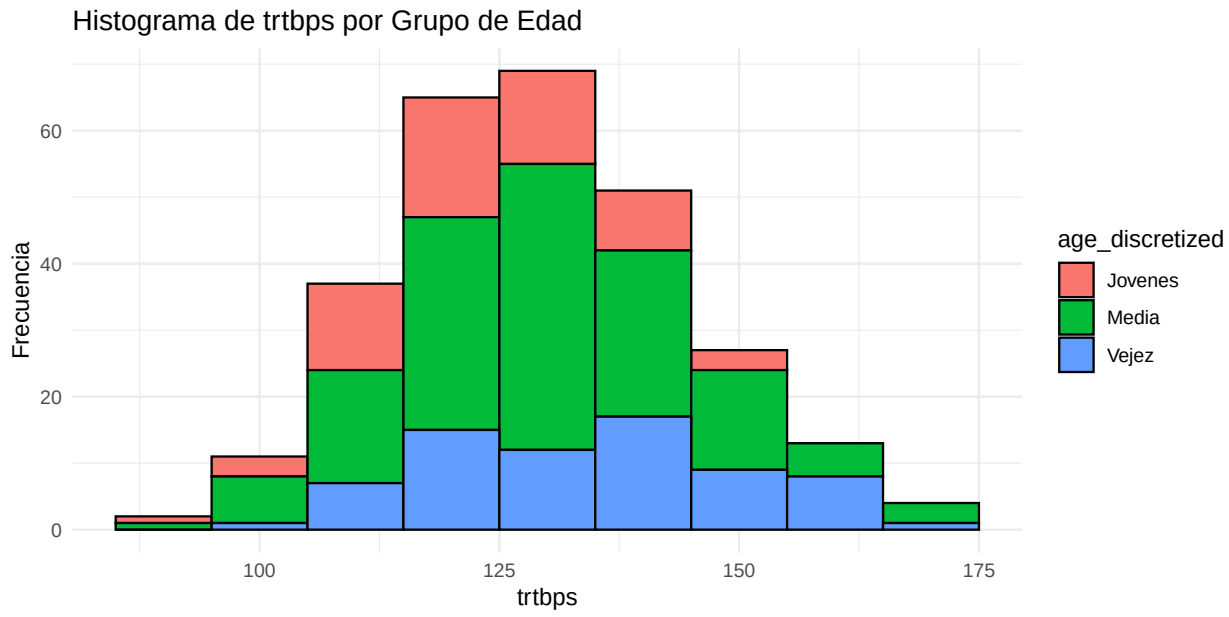


El valor crítico marcado para los grados de libertad de nuestros datos indican el nivel a partir del cual se considera que los resultados son estadísticamente significativos. En este caso observamos como el valor observado es inferior al valor crítico, por lo que, los resultados del test son no significativos.

TRTBPS

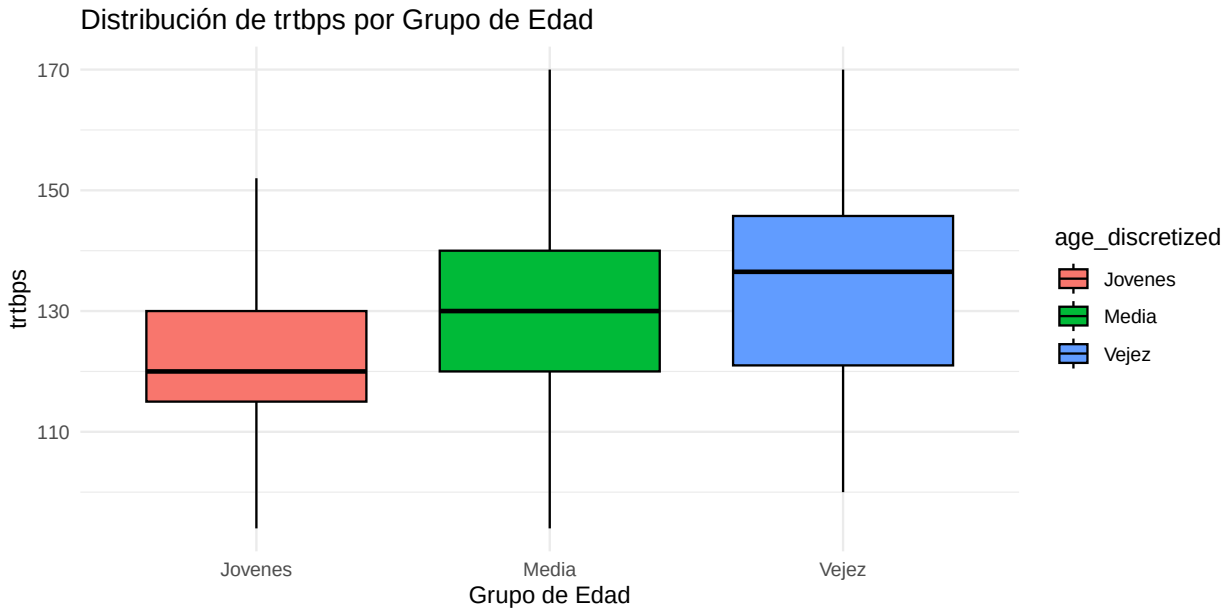
Para compara los valores de esta variable en los grupos de edad, hemos empleado Kruskal-Wallis, que afirma que las medianas de los datos en los diferentes grupos son iguales. Por lo que, el primer paso será visualizar la distribución de nuestros datos en cada uno de nuestros grupos.

Generamos el histograma de la variable `trtbps` en cada grupo de edad:



Se observa que el patrón entre los jóvenes y el grupo de edad media es similar y el de vejez es diferente a ambos.

A continuación realizaremos el boxplot de cada grupo, ya que esta visualización es la que nos permitirá identificar claramente si existen diferencias entre la mediana de los grupos.



Observamos que existen diferencias de la media de trtbps para cada grupo de edad. Además, para vejez, observamos una caja más ensanchada, lo que significa que la variabilidad de los datos es mayor en ese grupo.

Calculamos la media para cada grupo y la mostramos:

```
##  Jovenes    Media    Vejez
## 123.3770 130.0135 135.0857
```

Estos resultados concuerdan con el test de Kruskall-Wallis donde se concluía que el valor de estas variables variaba en función del grupo de edad.

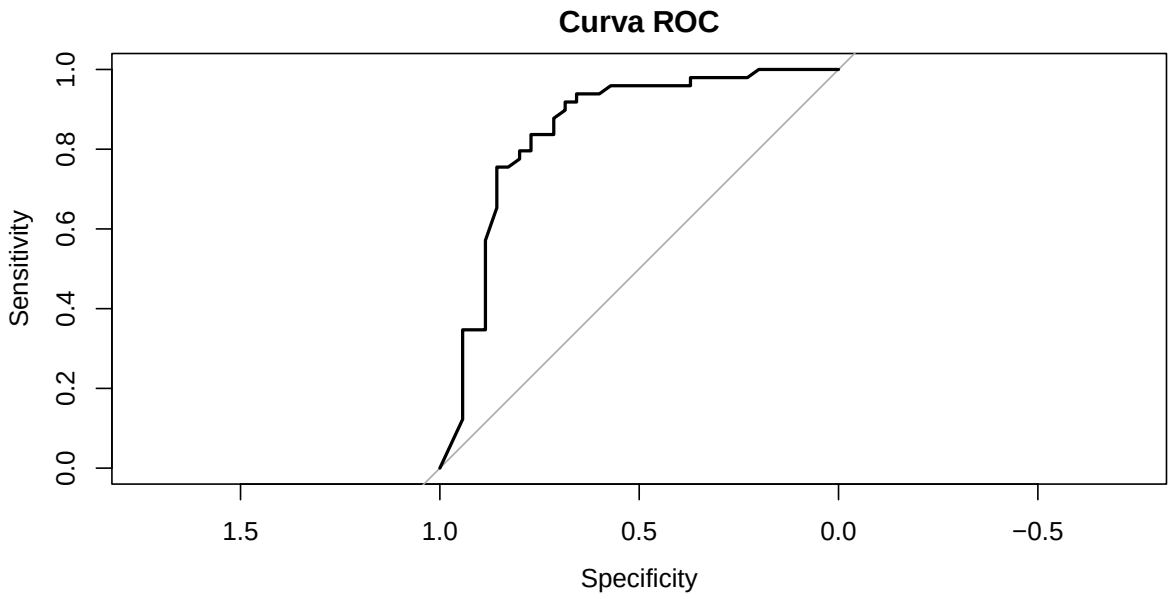
REGRESIÓN LOGÍSTICA

A partir del modelo obtenido en el apartado anterior predeciremos la variable output para el conjunto de test, y compararemos con los resultados reales. Visualizaremos la matriz de confusión.

```
##
## predicted         0         1
##     FALSE 0.8064516 0.1935484
##     TRUE  0.1886792 0.8113208
```

Por último, calcularemos la curva ROC y evaluamos el modelo con el coeficiente AUC, que indica el valor del área bajo la curva.

Visualizamos la curva:



Visualizamos el valor del coeficiente bajo la curva:

```
##
## Call:
## roc.default(response = test$output, predictor = predicted_probs)
##
## Data: predicted_probs in 35 controls (test$output 0) < 49 cases (test$output 1).
## Area under the curve: 0.8472
```

Del resultado del modelo podemos concluir que:

La ecuación del modelo ajustado es logit(p) = 1.2937 + 1.7425 * cp1 + 2.5354 * cp2 + 1.8978 * cp3 - 1.1598 * oldpeak - 2.6510 * caa1 - 2.1125 * caa2 - 3.3442 * caa3.

- Las variables que presentan significancia respecto a la variable de salida (infarto) son cp, oldpeak y caa.
- Los coeficientes de la ecuación del modelo indican que, por cada unidad que aumenta la variable correspondiente, el log-odds de la variable de salida aumenta en el valor del coeficiente asociado.
- La matriz de confusión indica que el modelo predice correctamente el 80.65% de los valores negativos (no infarto) y el 81.13% de los valores positivos (infarto).
- El valor de AUC de **0.8472** indica un buen rendimiento del modelo en la clasificación de los datos.

6. Resolución del problema

Los resultados obtenidos son concluyentes y brindan respuestas satisfactorias a las preguntas planteadas. Nuestro modelo logístico ha demostrado ser efectivo al predecir la probabilidad de sufrir un ataque al corazón. Al explorar los datos, hemos descubierto correlaciones y diferencias significativas entre diferentes grupos de pacientes. Por ejemplo, no se encontraron diferencias significativas en cuanto al tipo de dolor en el pecho y el sexo dentro de la población que sufre un infarto. Esto sugiere que estas variables no tienen un impacto

significativo en la probabilidad de padecer un ataque al corazón en la muestra analizada. Sin embargo, hemos observado que los niveles de presión sanguínea varían en función de la edad, lo que indica una relación entre la edad de los pacientes y este factor. Estos hallazgos sugieren que la edad puede ser un factor influyente en el riesgo de sufrir un infarto. Por último, nuestro modelo de regresión logística ha demostrado un buen rendimiento en la clasificación de los datos, con una precisión de alrededor del 80%.

7. Código

Además de en el archivo `TCVD-PRA2-HeartAttack.Rmd`, también habrá uno llamado `TCVD-PRA2-code.R` dentro de la misma carpeta `/code`

8. Vídeo

El vídeo esta colgado en VideoPEC del campus de la UOC y se puede acceder desde el siguiente **enlace**.

https://cv.uoc.edu/app/blogaula222/222_m2_851_01_448590/2023/06/04/practica-2-limpieza-y-analisis-de-datos/?ili=1

9. Contribuciones

Cuadro 2: Resumen de contribución en la práctica

Contribuciones	Firma
Investigación previa	V.M.G, M.E.M.M
Redacción de las respuestas	V.M.G, M.E.M.M
Desarrollo del código	V.M.G, M.E.M.M
Participación en el vídeo	V.M.G, M.E.M.M