

UOC - Tipología 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 junio 2023

Índice

1. Descripción del dataset	2
2. Integración y selección de variables	4
3. Limpieza de los datos	7
3.1. ¿Los datos contienen ceros o elementos vacíos?	8
3.2. Identifica y gestiona los valores extremos	9
4. Análisis de los datos	10
4.1. Selección de los grupos de datos que se quieren analizar/comparar	10
4.2. Comprobación de la normalidad y homogeneidad de la varianza	11
4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos	14
5. Representación de los resultados	17
6. Resolución del problema	27
7. Código	27
8. Vídeo	27

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:

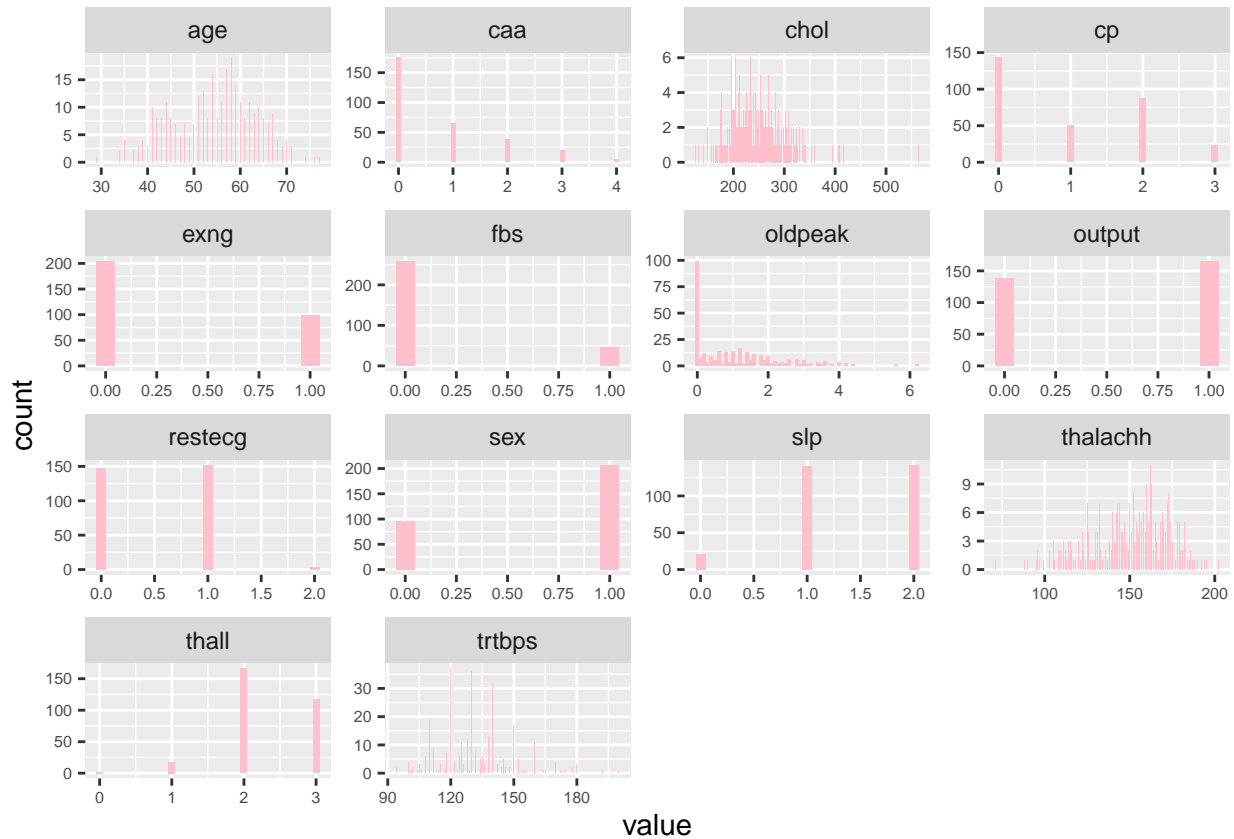
```
## 'data.frame': 303 obs. of 14 variables:
## $ age : int 63 37 41 56 57 57 56 44 52 57 ...
## $ sex : int 1 1 0 1 0 1 0 1 1 1 ...
## $ cp : int 3 2 1 1 0 0 1 1 2 2 ...
## $ trtbps : int 145 130 130 120 120 140 140 120 172 150 ...
## $ chol : int 233 250 204 236 354 192 294 263 199 168 ...
## $ fbs : 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 numérica discreta a excepción de **oldpeak** que se ha cargado como numérica continua.

A continuación realizaremos una visión general del dataset con `glimpse()`.

```
## Rows: 303
## Columns: 14
## $ age <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
## $ sex <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp <int> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
## $ trtbps <int> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
## $ chol <int> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ restecg <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
## $ thalachh <int> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exng <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0~
## $ oldpeak <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
## $ slp <int> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
## $ caa <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2~
## $ thall <int> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ output <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

Y una visión en general 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()`:

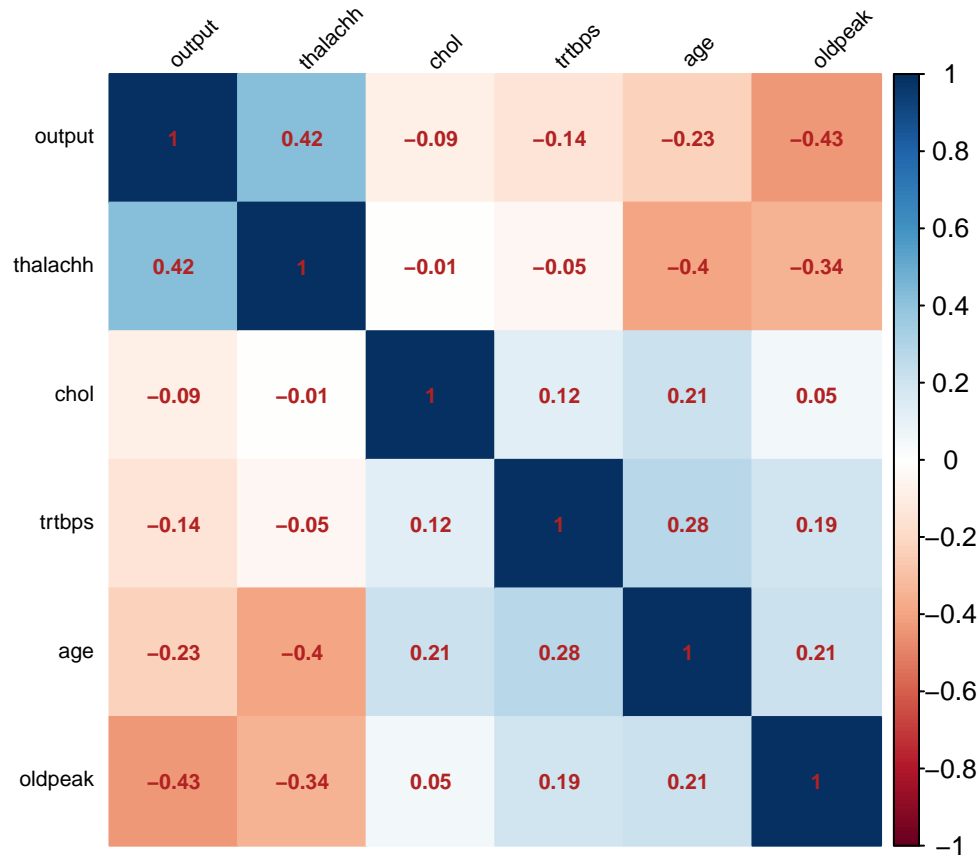
```
# Las convertimos a factor
heartAttack <- heartAttack %>%
  mutate(across(all_of(categorical_var), factor))
```

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 oxígeno 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 sería 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()`

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

```
# Creamos una lista vacía para almacenar los resultados de las pruebas de Fisher
fisher_results <- list()

# Iteramos sobre cada variable categórica
for (var in categorical_var) {
  fisher_result <- fisher.test(heartAttack[[var]], heartAttack$output)
  fisher_results[[var]] <- fisher_result
}
```

Visualizamos los resultados obtenidos del test de Fisher:

```
print(fisher_results)
```

```
## $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 categoricas 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.

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

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

Y ahora se va a ver un 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.

```
##      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
```

De la variable `caa` se tenían identificados valores de 0 a 3, **pero el valor máximo 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 dicotómicas `sex`, `fbs`, `exng` y `output`. Además existe la variable `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


```
colSums(heartAttack %>% select(all_of(numerical_var)) == 0)
```

```
##      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()`

```
colSums(is.na(heartAttack))
```

```
##      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

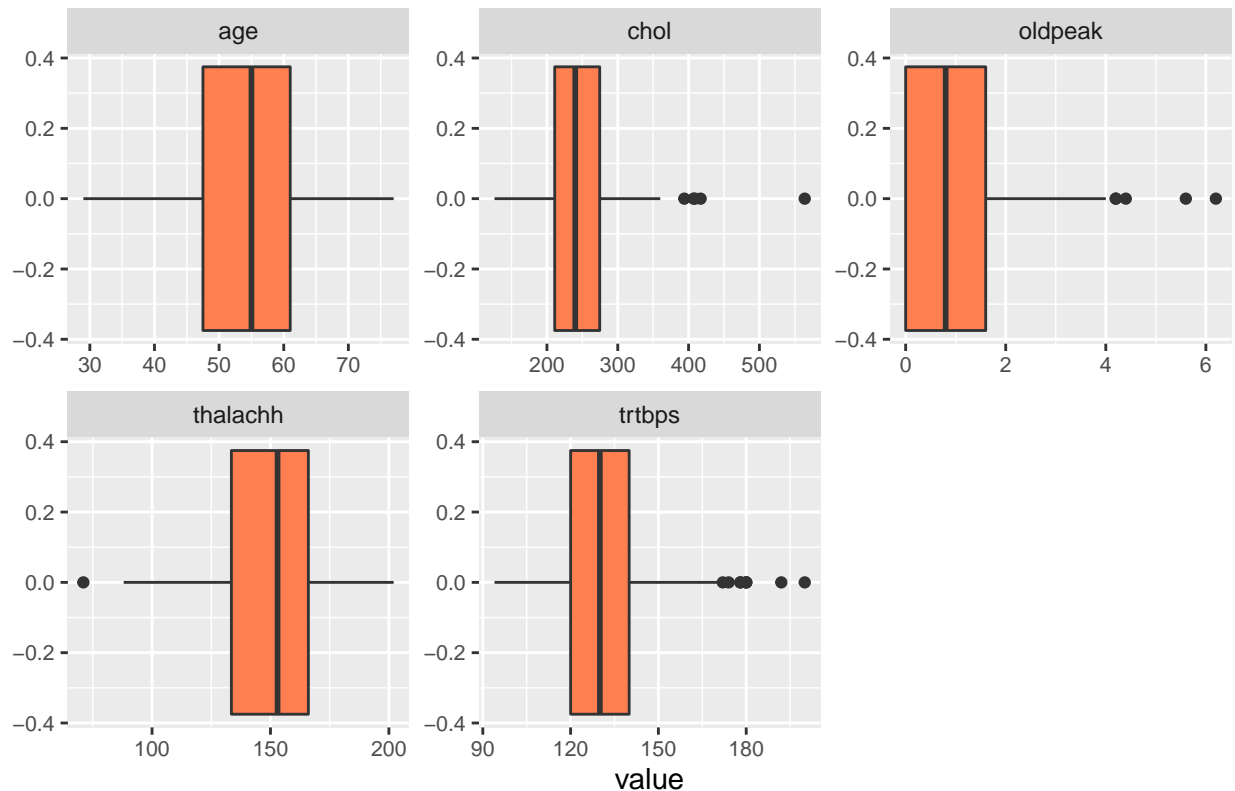
Primero en las variables numéricas vamos a comprobar cuantos valores atípicos hay de cada una

```
##      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.

Boxplot para buscar Outliers



También queremos ver si se va a perder mucha información a la hora de borrar registros por ello veremos cuantos registros tenemos antes 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 5ª 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:** Queremos conocer si la media de las variables numéricas **trtbps**, **chol** son las mismas 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.

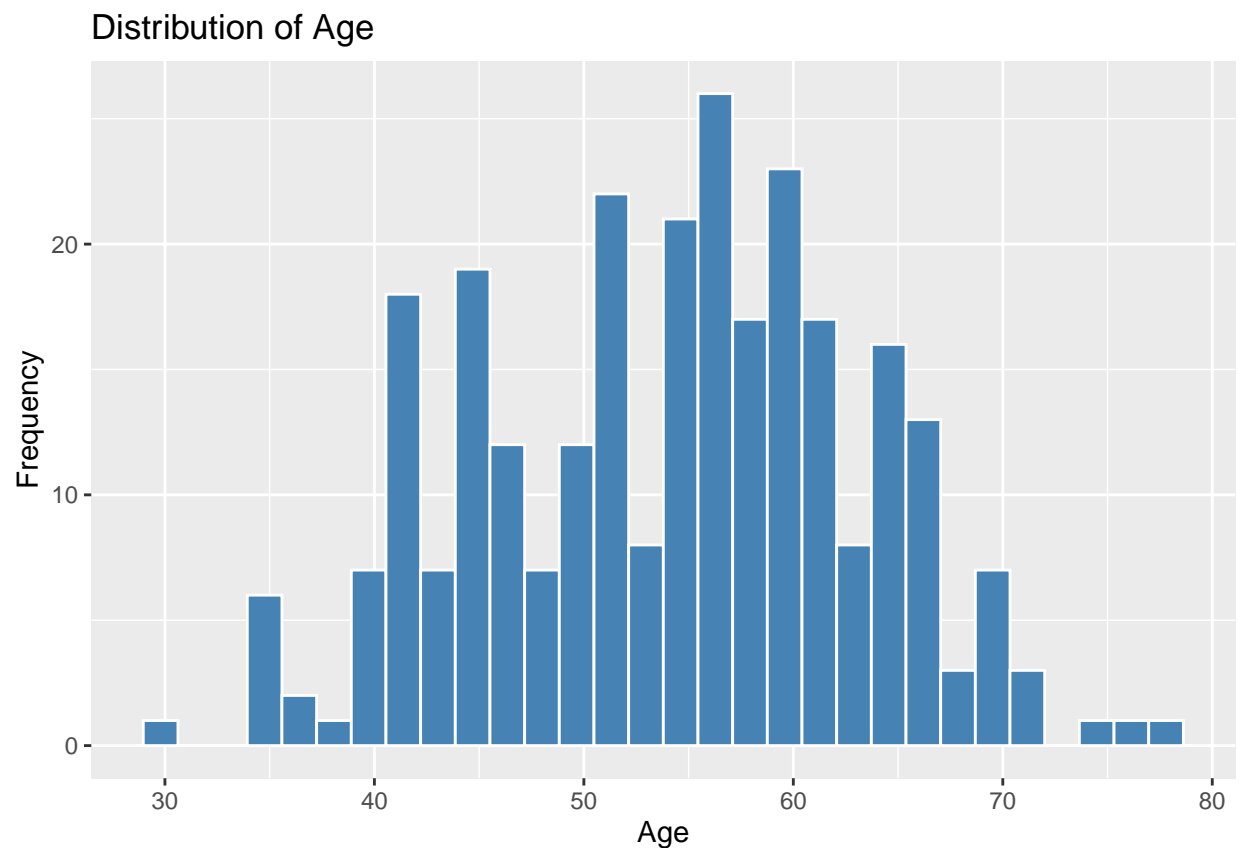
FIXME

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

```
# Calculamos los cuantiles
age_breaks <- c(0, 45, 60, 80)

# Discretizamos la variable age en grupos
heartAttack$age_discretized <-
  cut(
    heartAttack$age,
    breaks = age_breaks,
    labels = c("Jovenes", "Media", "Vejez"),
```

```

    include.lowest = TRUE
)

```

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

```

# Obtener el máximo y mínimo de la variable "age_discretized" en el conjunto de datos "heartAttack"
max_min_age <- tapply(heartAttack$age, heartAttack$age_discretized, range)

```

```

# Obtener el máximo y mínimo para el grupo "tercio1"
max_jovenes <- max_min_age[["Jovenes"]][2]
min_jovenes <- max_min_age[["Jovenes"]][1]

```

```

# Obtener el máximo y mínimo para el grupo "tercio2"
max_medio <- max_min_age[["Media"]][2]
min_medio <- max_min_age[["Media"]][1]

```

```

# Obtener el máximo y mínimo para el grupo "tercio3"
max_vejez <- max_min_age[["Vejez"]][2]
min_vejez <- max_min_age[["Vejez"]][1]

```

```

cat("Jovenes -", "Mínimo:", min_jovenes, "Máximo:", max_jovenes, "\n")

```

```

## Jovenes - Mínimo: 29 Máximo: 45

```

```

cat("Medio -", "Mínimo:", min_medio, "Máximo:", max_medio, "\n")

```

```

## Medio - Mínimo: 46 Máximo: 60

```

```

cat("Vejez -", "Mínimo:", min_vejez, "Máximo:", max_vejez, "\n")

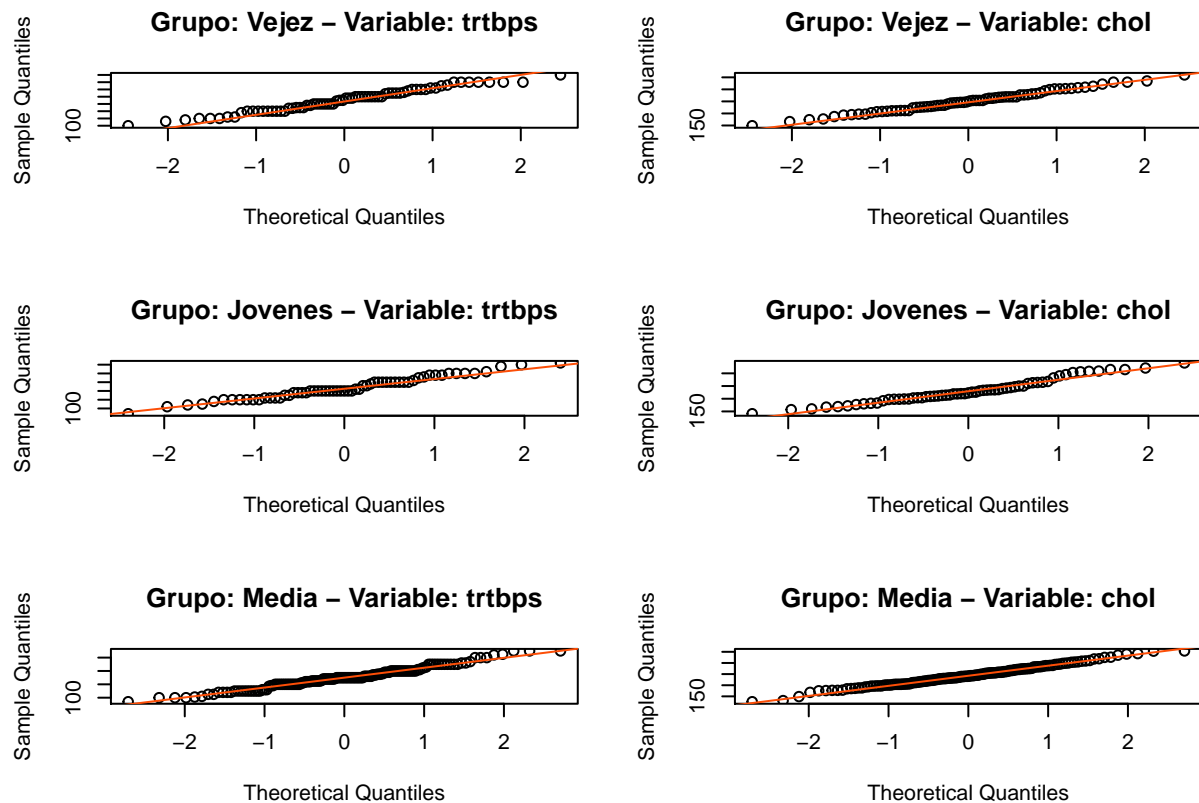
```

```

## Vejez - Mínimo: 61 Máximo: 77

```

Ahora comprobamos si las variables trtbps, chol presentan una distribución normal de manera visual, a través del gráfico Q-Q:



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

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

```
# Comprobamos la normalidad por grupo
group_names <- unique(heartAttack$age_discretized)

for (group in group_names) {
  group_data <- subset(heartAttack, age_discretized == group)

  shapiro_test_chol <- shapiro.test(group_data$chol)
  shapiro_test_trtbps <- shapiro.test(group_data$trtbps)

  cat("Grupo:", group, "\n")
  cat("El p-value de chol es:", shapiro_test_chol$p.value, "\n")
  cat("El p-value de trtbps es:", shapiro_test_trtbps$p.value, "\n")

  if (shapiro_test_chol$p.value >= 0.05) {
    cat("La variable 'colesterol' en el grupo", group, "sigue una distribución normal.\n")
  } else {
    cat("La variable 'colesterol' en el grupo", group, "no sigue una distribución normal.\n")
  }

  if (shapiro_test_trtbps$p.value >= 0.05) {
    cat("La variable 'trtbps' en el grupo", group, "sigue una distribución normal.\n")
  }
}
```

```

} else {
  cat("La variable 'trtbps' en el grupo", group, "no sigue una distribución normal.\n")
}
cat("\n")
}

```

```

## Grupo: Vejez
## El p-value de chol es: 0.9713229
## El p-value de trtbps es: 0.2293937
## La variable 'colesterol' en el grupo Vejez sigue una distribución normal.
## La variable 'trtbps' en el grupo Vejez sigue una distribución normal.
##
## Grupo: Jovenes
## El p-value de chol es: 0.09457119
## El p-value de trtbps es: 0.2786165
## La variable 'colesterol' en el grupo Jovenes sigue una distribución normal.
## La variable 'trtbps' en el grupo Jovenes sigue una distribución normal.
##
## Grupo: Media
## El p-value de chol es: 0.6880372
## El p-value de trtbps es: 0.03323032
## La variable 'colesterol' en el grupo Media sigue una distribución normal.
## 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 los grupos vejez y media, la variable `chol` y `trtbps` no presentan 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.

```
infarto <- subset(heartAttack, output== 1)
```

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

```
tabla <- table(infarto$cp, infarto$sex)
tabla
```

```

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

```

Realizamos la prueba chi-cuadrado:

```
# Prueba de chi-cuadrado
chi_square <- chisq.test(tabla)
```

```
# Imprimimos los resultados
print(chi_square)
```

```
##
## 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 las variables **trtbps** y **chol** entre los grupos de edad **Jovenes**, **Media y Vejez**, y como sabemos que no se cumple en supuesto de normalidad de los datos en cada grupo, aplicaremos Kruskal_Wallis.

Nuestra hipótesis son:

- Hipótesis nula (H₀): No hay diferencias significativas de la media del valor trtbps y chol entre los diferentes grupos de edad.
- Hipótesis alternativa (H₁): Hay diferencias significativas de la media del valor trtbps y chol entre los diferentes grupos de edad.

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

```
kruskal.test(trtbps~age_discretized, data=heartAttack)
```

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

Realizamos el test de Kruskal-Wallis para el colesterol y los diferentes grupos de edad:

```
kruskal.test(chol~age_discretized, data=heartAttack)
```

```
##
## Kruskal-Wallis rank sum test
##
## data:  chol by age_discretized
## Kruskal-Wallis chi-squared = 7.0282, df = 2, p-value = 0.02977
```

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

REGRESIÓN LOGÍSTICA

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.

```
#Utilizaremos el 70% para train y el 30% para test
set.seed(23)
```

```
train_index <- sample(1:nrow(heartAttack), nrow(heartAttack) * 0.7)
train <- heartAttack[train_index, ]
test <- heartAttack[-train_index, ]
```

Ajustamos el modelo de regresión logística.

```
model <- glm(output ~ ., data = train, family = binomial)
```

Visualizamos el resultado del modelo:

```
# Obtenemos los resultados del modelo
summary(model)
```

```
##
## 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, restecg, oldpeak y caa.

Por lo tanto, volvemos a realizar un modelo con solo esas variables.

# Ajustamos un nuevo modelo con las variables significativas
model_2 <- glm(output ~ cp + restecg + oldpeak + caa, data = train, family = binomial)

summary(model_2)

##
## Call:
## glm(formula = output ~ cp + restecg + oldpeak + caa, family = binomial,
##      data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.5176  -0.5029   0.2028   0.5549   2.1798
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.2108     0.4524   2.676 0.007440 **
## cp1             1.7363     0.6200   2.800 0.005103 **
## cp2             2.5081     0.5748   4.363 1.28e-05 ***
## cp3             1.9155     0.7861   2.437 0.014818 *
## restecg1        0.1551     0.4213   0.368 0.712718
## restecg2       13.1265    1455.3977   0.009 0.992804
## oldpeak        -1.1630     0.2485  -4.681 2.86e-06 ***
## caa1            -2.6255     0.5401  -4.861 1.17e-06 ***
## caa2            -2.1105     0.6769  -3.118 0.001820 **
## caa3            -3.3078     0.9909  -3.338 0.000843 ***
## ---
## 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: 145.17  on 185  degrees of freedom
## AIC: 165.17
##
## Number of Fisher Scoring iterations: 14
```

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.

```
# Representación de la variable cp en función de los grupos
infarto_porcentaje <- infarto %>%
  group_by(sex, cp) %>%
  summarize(n = n()) %>%
```

```
mutate(percentage = n / sum(n))

## `summarise()` has grouped output by 'sex'. You can override using the `.groups` argument.
# Mapeo de etiquetas para las categorías de cp y sex
etiquetas_cp <- c("Angina típica", "Angina atípica", "Dolor no anginal", "Asintomático")
etiquetas_sex <- c("Femenino", "Masculino")

# Crear un gráfico de líneas
ggplot(data = infarto_porcentaje, aes(x = cp, y = percentage, color = sex, group = sex)) +
  geom_line() +
  labs(x = "cp", y = "Porcentaje de observaciones (%)") +
  ggtitle("Porcentaje de observaciones por cp y sexo") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal() +
  scale_x_discrete(labels = etiquetas_cp) +
  scale_color_discrete(labels = etiquetas_sex)
```



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, habiéndose la mayor diferencia entre sexos en el Dolor no Anginal, donde las mujeres lo experimentan en más 10% por encima de los hombres.

El test Chi-cuadrado no ha evaluado si existe una diferencia significativa entre estas dos variables categóricas, y para ello evalúa la diferencia entre las frecuencias observadas y las frecuencias esperadas, en el supuesto nulo de independencia. Ahora vamos a graficar los valores observados y los esperados.

```
library(ggplot2)
```

```

# Obtener los valores observados y esperados del test chi-cuadrado
observados <- chi_square$observed
esperados <- chi_square$expected

# Categorías para las variables sex y cp
categorias_sex <- c("Femenino", "Masculino")
categorias_cp <- c("Angina típica", "Angina atípica", "Dolor no anginal", "Asintomático")

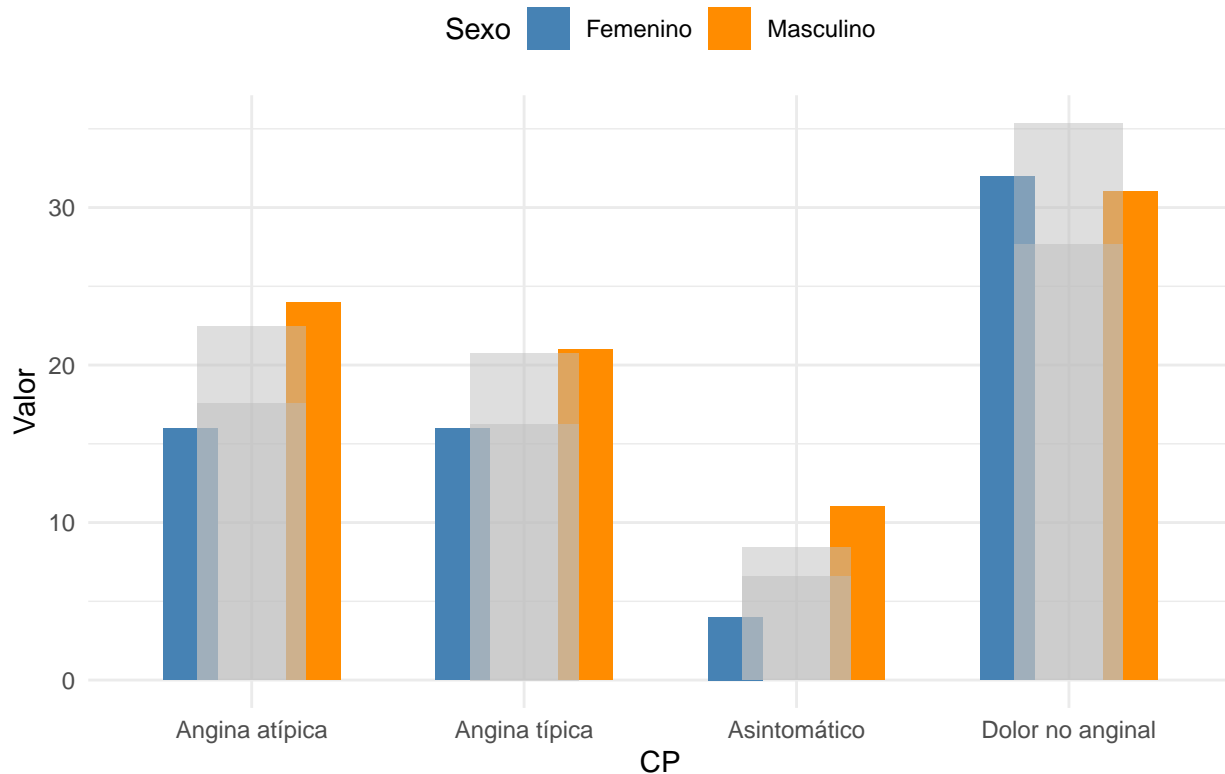
# Crear un dataframe con los valores observados y esperados
datos <- data.frame(cp = rep(categorias_cp, length(categorias_sex)),
                    sexo = rep(categorias_sex, each = length(categorias_cp)),
                    observados = as.vector(observados),
                    esperados = as.vector(esperados))

# Definir los colores para cada categoría de sexo
colores <- c("steelblue", "darkorange")

# Graficar los datos
ggplot(data = datos, aes(x = cp, fill = sexo)) +
  geom_bar(aes(y = observados), position = position_dodge(width = 0.9), stat = "identity", width = 0.4)
  geom_bar(aes(y = esperados), position = position_dodge(width = 0.9), fill = "gray", alpha = 0.5, stat = "identity", width = 0.4)
  labs(x = "CP", y = "Valor", fill = "Sexo") +
  ggtitle("Valores Observados y Esperados por CP y Sexo") +
  scale_fill_manual(values = colores) +
  theme_minimal() +
  theme(legend.position = "top")

```

Valores Observados y Esperados por CP y Sexo



Visualizamos la tabla de valores:

datos

```
##          cp      sexo observados esperados
## 1  Angina típica Femenino         16 16.232258
## 2  Angina atípica Femenino         16 17.548387
## 3 Dolor no anginal Femenino         32 27.638710
## 4  Asintomático Femenino           4  6.580645
## 5  Angina típica Masculino          21 20.767742
## 6  Angina atípica Masculino          24 22.451613
## 7 Dolor no anginal Masculino          31 35.361290
## 8  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.

```
# Calcular los grados de libertad
grados_libertad <- length(unique(infarto$sex)) * (length(unique(infarto$cp)) - 1)

# Calcular los valores para la distribución chi-cuadrado
x <- seq(0, 20, 0.1) # Valores en el rango de 0 a 20 (ajusta según tus necesidades)
y <- dchisq(x, df = grados_libertad) # Función de densidad chi-cuadrado

# Graficar la distribución chi-cuadrado
plot(x, y, type = "l", lwd = 2,
     xlab = "Valor chi-cuadrado", ylab = "Densidad",
     main = "Distribución Chi-Cuadrado")
```

```

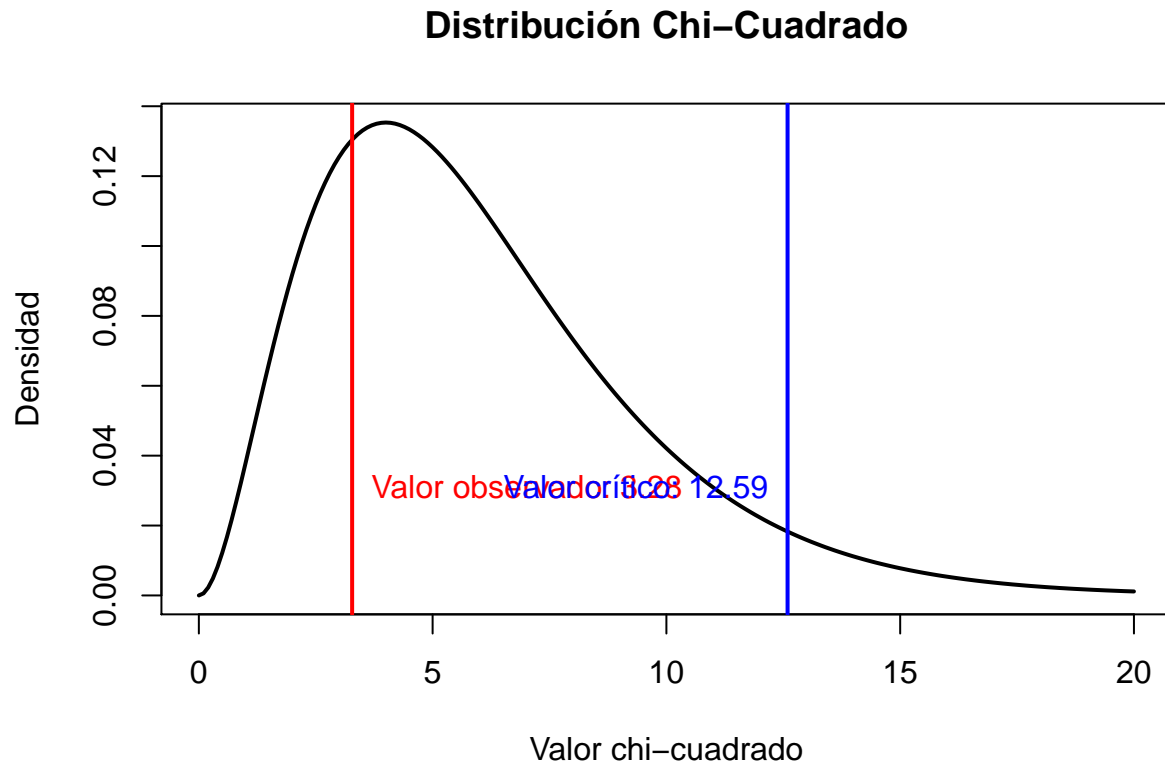
# Obtener el valor observado del estadístico chi-cuadrado
valor_observado <- chi_square$statistic

# Agregar línea vertical para el valor observado chi-cuadrado
abline(v = valor_observado, col = "red", lwd = 2)
text(valor_observado, 0.03, paste("Valor observado:", round(valor_observado, 2)),
     pos = 4, col = "red")

# Calcular el valor crítico
nivel_significancia <- 0.05
valor_critico <- qchisq(1 - nivel_significancia, df = grados_libertad)

# Agregar línea vertical para el valor crítico
abline(v = valor_critico, col = "blue", lwd = 2)
text(valor_critico, 0.03, paste("Valor crítico:", round(valor_critico, 2)),
     pos = 2, col = "blue")

```



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 tanto los resultados del test son no significativos.

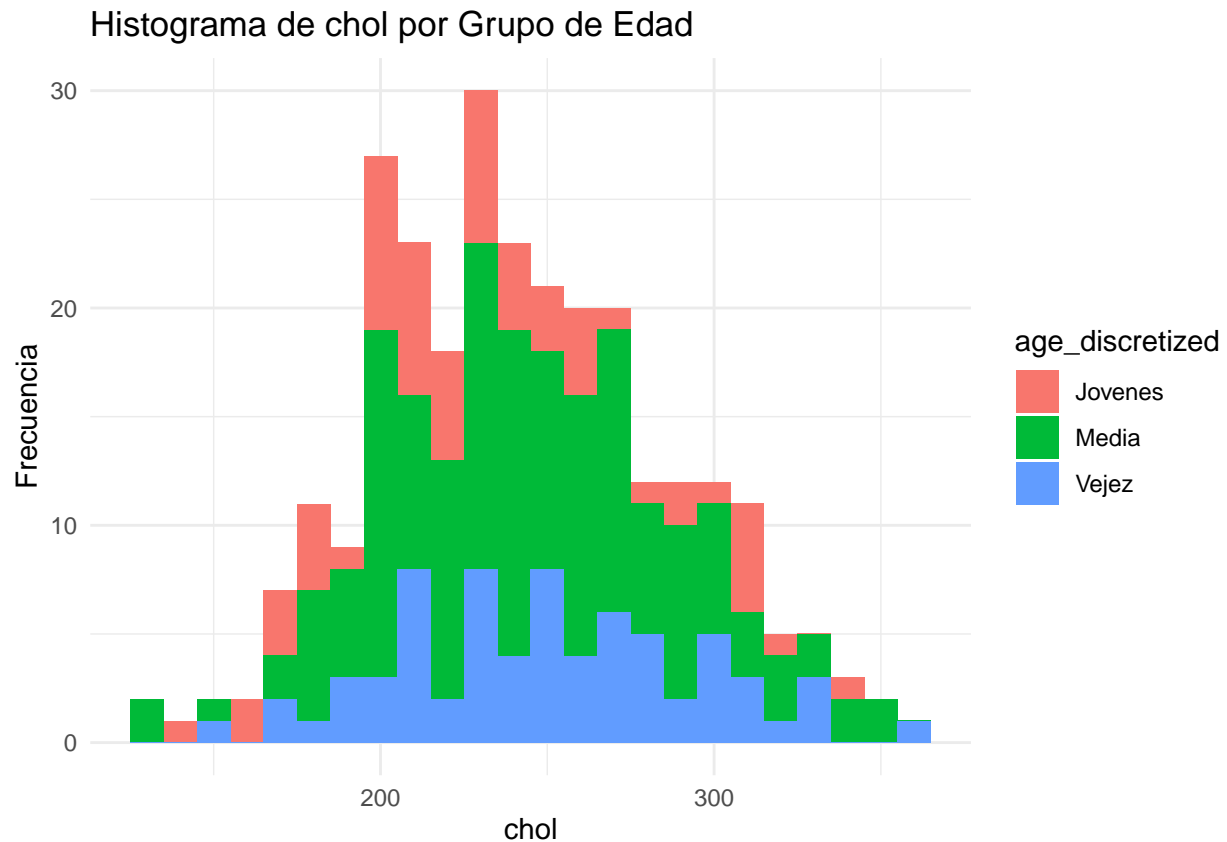
CHOL y TRTBPS

Para comparar estas dos variables 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 chol en cada grupo de edad:

```
heartAttack_filtered <- heartAttack[, c("chol", "age_discretized")]

ggplot(heartAttack_filtered, aes(x = chol, fill = age_discretized)) +
  geom_histogram(binwidth = 10, position = "stack") +
  labs(x = "chol", y = "Frecuencia") +
  ggtitle("Histograma de chol por Grupo de Edad") +
  theme_minimal()
```

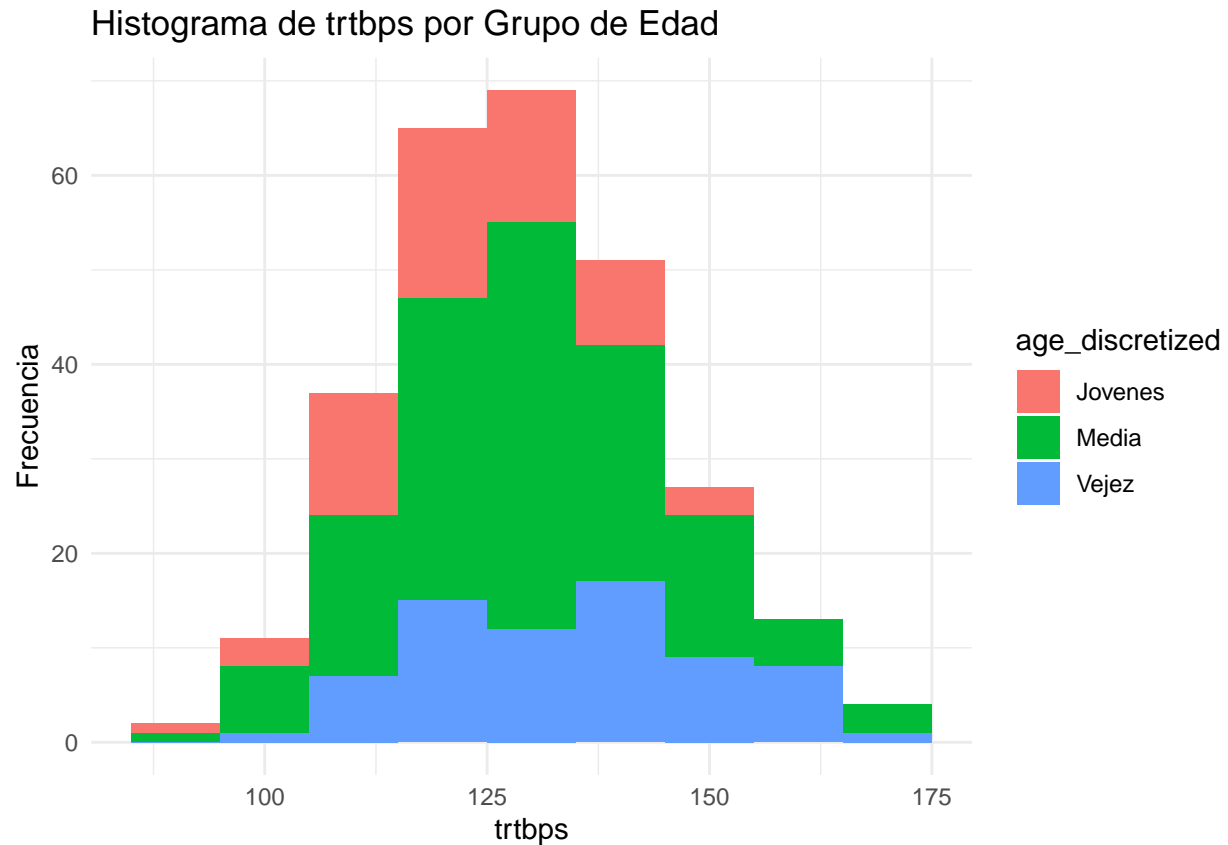


Observamos que para media y jóvenes la distribución es similar, pero para vejez parece otro patrón.

Realizamos el mismo paso para la variable trtbps:

```
heartAttack_filtered <- heartAttack[, c("trtbps", "age_discretized")]

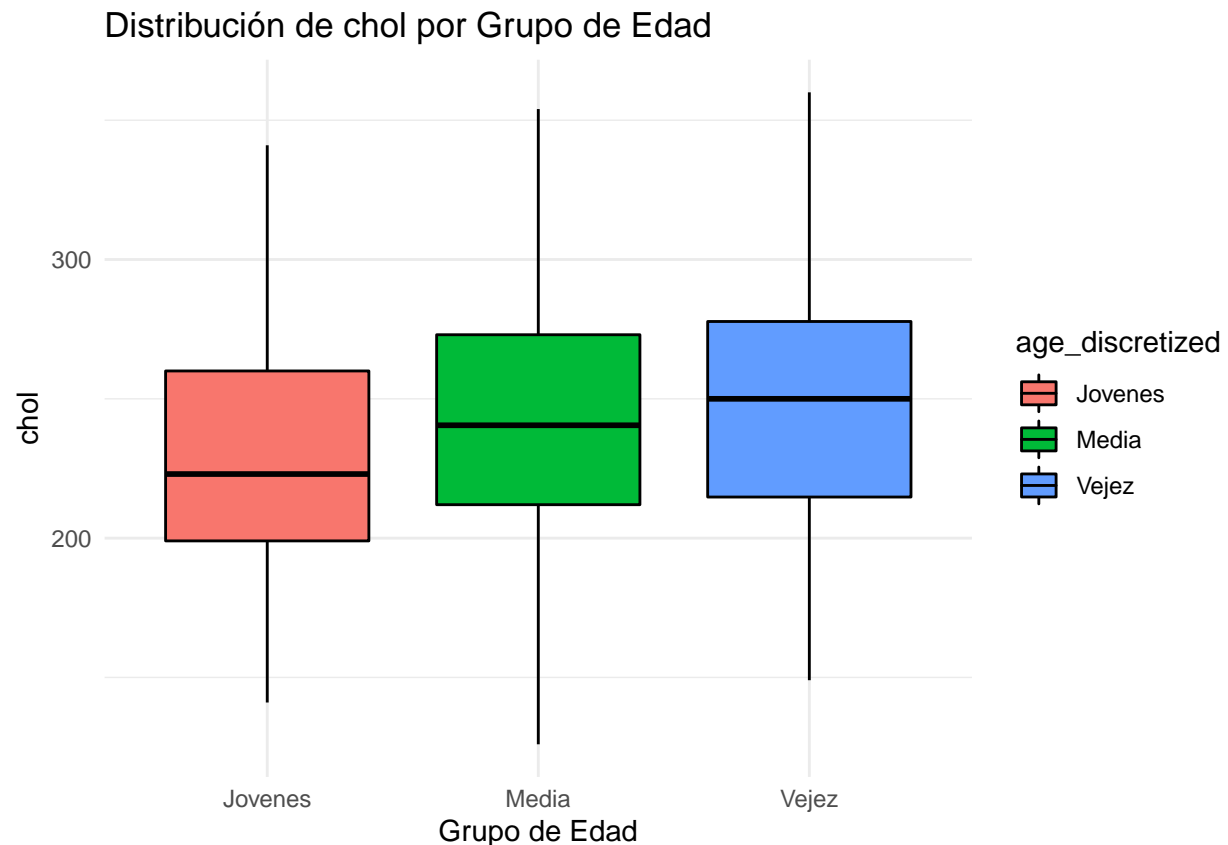
ggplot(heartAttack_filtered, aes(x = trtbps, fill = age_discretized)) +
  geom_histogram(binwidth = 10, position = "stack") +
  labs(x = "trtbps", y = "Frecuencia") +
  ggtitle("Histograma de trtbps por Grupo de Edad") +
  theme_minimal()
```



Los resultados son similares a los obtenidos con la variable chol, donde parece que el patron entre Jovenes y 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.

```
# Crear el gráfico de cajas
ggplot(heartAttack, aes(x = age_discretized, y = chol, fill = age_discretized)) +
  geom_boxplot(color = "black") +
  labs(x = "Grupo de Edad", y = "chol") +
  ggtitle("Distribución de chol por Grupo de Edad") +
  theme_minimal()
```



Para la variable chol observamos que existen diferencias en su media para cada grupo. Visualizamos este valor para cada grupo:

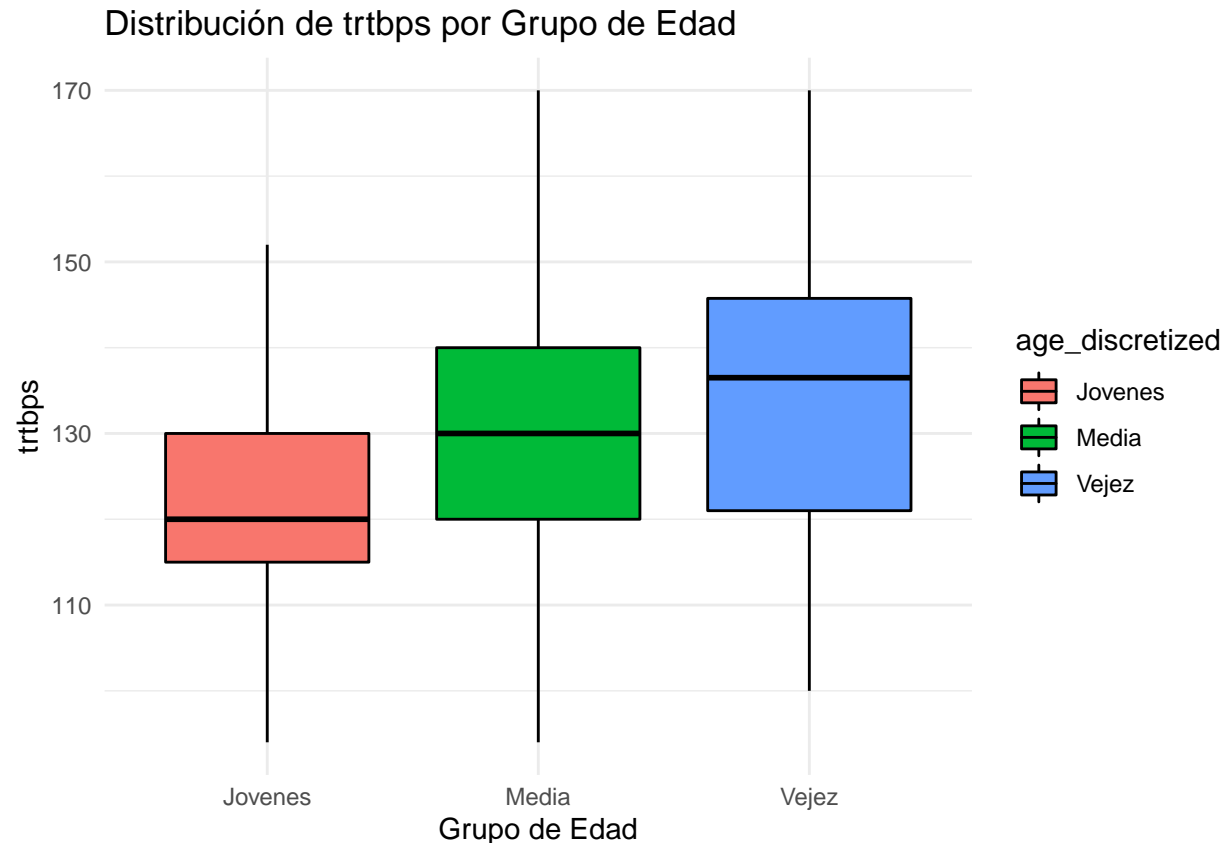
```
# Calcular la media por grupo
means <- tapply(heartAttack$chol, heartAttack$age_discretized, mean)

# Mostrar las medias por grupo
print(means)
```

```
## Jovenes Media Vejez
## 231.0984 243.7230 250.6000
```

Realizamos el boxplot para la variable trtbps:

```
# Crear el gráfico de cajas
ggplot(heartAttack, aes(x = age_discretized, y = trtbps, fill = age_discretized)) +
  geom_boxplot(color = "black") +
  labs(x = "Grupo de Edad", y = "trtbps") +
  ggtitle("Distribución de trtbps por Grupo de Edad") +
  theme_minimal()
```

Observamos que se cumple el mismo patrón que en el caso anterior, y para vejez, además, 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:

```
# Calcular la media por grupo
means <- tapply(heartAttack$trtbps, heartAttack$age_discretized, mean)

# Mostrar las medias por grupo
print(means)
```

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

Estos resultados concuerdan con el test de Kruskal-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 prediciremos la variable output para el conjunto de test, y compararemos con los resultados reales. Visualizaremos la matriz de confusión.

```
#Predicimos
predicted <- predict(model_2, newdata = test, type = "response") > 0.5

#realizamos la matriz de confusion
confusion_matrix <- table(predicted, test$output)

# Normalizamos la matriz de confusión
```

```
normalized_confusion_matrix <- prop.table(confusion_matrix, margin = 1)
```

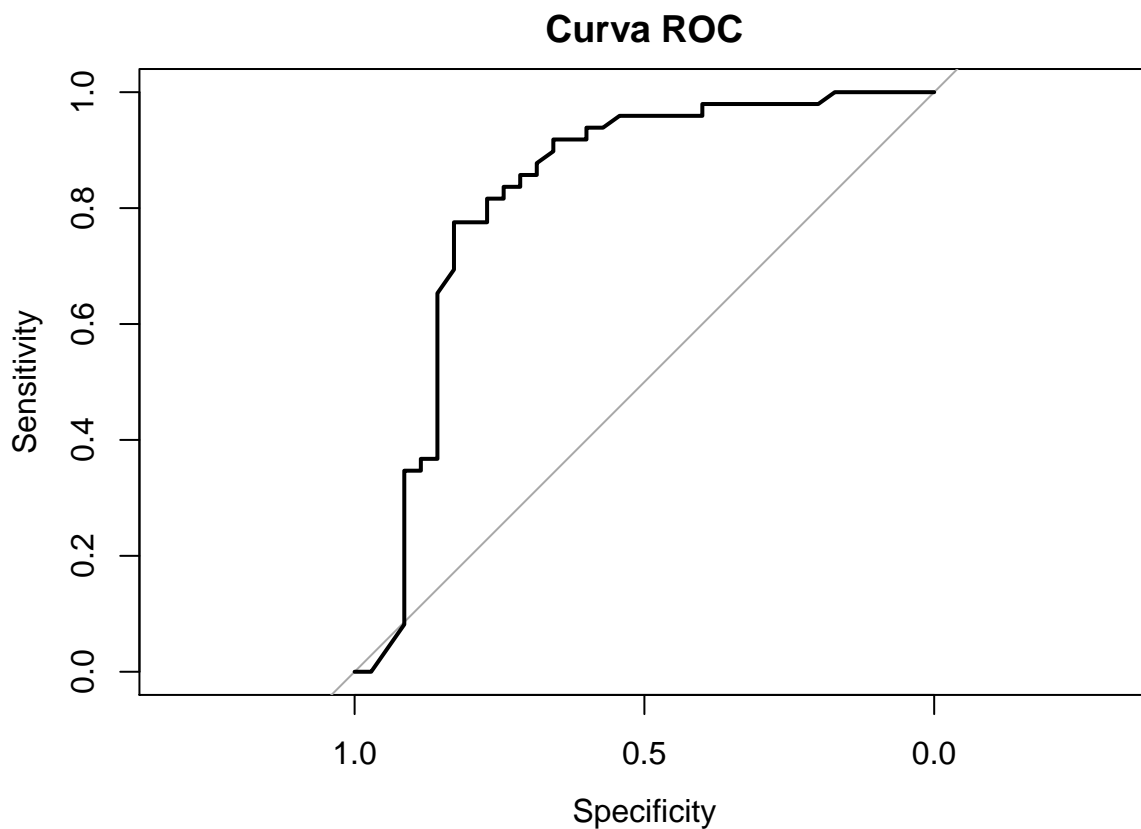
```
# Visualizamos la matriz de confusión normalizada
print(normalized_confusion_matrix)
```

```
##
## predicted      0      1
##   FALSE 0.8000000 0.2000000
##   TRUE  0.2037037 0.7962963
```

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

Visualizamos la curva:

```
plot(roc_obj, main = "Curva ROC")
```



Visualizamos el valor del coeficiente bajo la curva:

```
auc
```

```
##
## 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: 0.8224
```

Del resultado del modelo podemos concluir que:

La ecuación del modelo ajustado es $\text{logit}(p) = 0.4046 + 1.5603 * \text{cp1} + 2.1369 * \text{cp2} + 1.8318 * \text{cp3} + 0.6486 * \text{restecg1} - 13.0644 * \text{restecg2} - 0.9669 * \text{oldpeak} - 0.6658 * \text{caa}$

- Las variables que presentan significancia respecto la variable de salida (infarto) son cp, restecg, oldpeak y caa
- Los coeficientes de la ecuación del modelo indican por cada unidad que aumenta es variable, cuanto aumenta el log-odds de la variable output.
- La matriz de confusión normalizada indica que el modelo predice el 84,21 % de los valores falsos correctamente, con una tasa de error del 15,79 %. Igualmente el modelo predice correctamente el 77,36 % de los casos de infarto, con una tasa de error del 22,64 %.
- El valor de AUC de 0,8769 indica un buen rendimiento del modelo.

6. Resolución del problema

FIXME

7. Código

FIXME

8. Vídeo