Limpieza y Análisis de Datos

Lucas Gómez Torres y Joan Amengual Mesquida

13 de enero, 2023

Índice General

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

Actualmente cada vez sufren más personas ataques al corazón originados por diferentes factores como pueden ser el exceso de colesterol, el nivel de azúcar en la sangre, el consumo de tabaco, la presión arterial, la obesidad, la edad o la falta de ejercicio, entre muchos otros más, que pueden dar lugar a un daño permanente en el corazón como la insuficiencia cardiaca o a la muerte.

Por ello, los ataques al corazón son un problema muy grave que hay que intentar prevenir, analizando las diferentes variables que pueden influir a la hora de que una persona sufra un ataque al corazón o no, pudiendo responder a preguntas como por ejemplo:

- ¿Los hombres son más probables a sufrir un ataque?
- ¿El nivel de azúcar en sangre es determinante para que una persona pueda padecer un ataque?
- ¿Las personas mayores tienen más probabilidad de sufrir un ataque?
- ¿Qué factor es el más influye en un ataque?

El conjunto de datos está dividido en dos subconjuntos de datos:

- *heart.csv*: contiene toda la información sobre los pacientes, incluyendo si finalmente sufrieron un ataque al corazón o no. Tiene 303 observaciones y 14 atributos. De estos 14 atributos, 13 son variables independientes y 1 la variable dependiente (nuestra variable objetivo que servirá para construir un modelo de aprendizaje supervisado que nos permita predecir si un paciente tendrá un ataque al corazón o no). A continuación, se describen todos los atributos de este dataset:
  - **age**: Variable de tipo numérica. Determina la edad de la persona.
  - **sex**: Variable de tipo numérica. Refleja el género de la persona *(1 = masculino, 0 = femenino)*.
  - **cp**: Variable de tipo numérica. Identifica el tipo de dolor en el pecho *(0 = angina típica, 1 = angina atípica, 2 = dolor no anginoso, 3 = asintomático)*.
  - **trtbps**: Variable de tipo numérica. Indica la presión arterial en reposo en mg/dl.
  - **chol**: Variable de tipo numérica. Hace referencia al nivel de colesterol en mg/dl.
  - **fbs**: Variable de tipo numérica. Indica si el nivel de azúcar en sangre en ayunas es mayor a 120 mg/dl *(1 = verdadero, 0 = falso)*.
  - **restecg**: Variable de tipo numérica. Muestra los resultados electrocardiográficos en reposo *(0 = normal, 1 = anomalía de onda ST-T (inversiones de onda T y/o elevación o depresión ST de > 0,05 mV), 2 = hipertrofia ventricular izquierda probable o definida por los criterios de Estes)*.
  - **thalachh**: Variable de tipo numérica. Determina la frecuencia cardiaca máxima alcanzada.
  - **exng:**: Variable de tipo numérica. Indica si la angina ha sido inducida por el ejercicio *(1 = sí, 0 = no)*.
  - **oldpeak**: Variable de tipo numérica. Señala la depresión ST inducida por el ejercicio en relación con el descanso.
  - **slp**: Variable de tipo numérica. Muestra la pendiente del segmento ST de ejercicio máximo *(0 = inclinación hacia abajo, 1 = plano, 2 = inclinación hacia arriba)*.
  - **caa**: Variable de tipo numérica. Indica el número de buques principales *(0, 1, 2, 3)*.

  - **thall**: Variable de tipo numérica. Señala el ratio de un trastorno sanguíneo llamado talasemia *(0 = no tiene, 1 = defecto fijo (sin flujo sanguíneo en alguna parte del corazón), 2 = flujo sanguíneo normal, 3 = defecto reversible (se observa un flujo sanguíneo, pero no es normal))*.
  - **output**: Variable de tipo numérica. Indica si el paciente sufre un ataque al corazón o no (0 = No, 1 = Sí). Se trata de la variable objetivo o dependiente que pretenderemos predecir.

- *o2Saturation.csv*: contiene 3585 observaciones sobre los niveles de oxígeno en la sangre de distintos pacientes y solo tiene 1 atributo.

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.

En este apartado se van a cargar ambos conjuntos de datos, para decidir si se van a unificar ambos o no, o si nos vamos a centrar en unos pasajeros concretos limitando el número de registros o de características con el fin de reducir el dataset. Además, en el dataset de *heart.csv* se van a renombrar los atributos para que se entiendan mejor y sean más intuitivos a la hora de utilizarlos más adelante.

```
# Se carga el dataset
heart_data <- read.csv("heart.csv", header = TRUE)

# Modificamos los nombres de las variables para que sean más intuitivos
colnames(heart_data) <- c("age","sex","chest_pain_type","resting_blood_pressure",
                          "cholesterol",    "fasting_blood_sugar","rest_ecg_type",
                          "max_heart_rate_achieved","exercise_induced_angina",
                          "st_depression", "st_slope_type", "num_major_vessels",
                          "thalassemia_type","heart_attack")

# Dimensión del dataset
dim(heart_data)
```

```
## [1] 303  14
```

```
# Se carga el dataset
O2_saturation <- read.csv("o2Saturation.csv", header = TRUE)

# Dimensión del dataset
dim(O2_saturation)
```

```
## [1] 3585    1
```

Podemos observar que ambos conjuntos de datos tienen dimensiones diferentes. El que contiene los niveles de oxígeno en la sangre consta de 3.585 observaciones, es decir, diferentes niveles de oxígeno para 3.585 pacientes, en cambio, el otro, contiene información sobre 303 pacientes y 14 características distintas. Como ya tenemos suficientes características en el dataset de *heart.csv* con las que poder realizar un estudio detallado y completo a las preguntas que hemos planteado al principio, se va a optar por descartar el otro conjunto y perder este atributo adicional de los pacientes.

En el caso de haber querido unificarlos y por lo tanto añadir otro atributo al dataset de *heart.csv* (saturación de oxígeno), se podría haber utilizado la función *merge* permitiéndonos fusionarlos de forma horizontal. Posteriormente, se podría comprobar que no existen inconsistencias ni duplicidades en los registros con la

función *duplicated* o *unique*. No obstante, no existe un identificador único para cada uno de los pacientes como podría ser un id o un nombre, por lo que suponemos que podría haber dos pacientes con los mismos valores de atributos. Asimismo comprobaremos si hay muchos registros duplicados con el fin de que no pueda afectar significamente en los análisis posteriores.

```
# Comprobamos si existen registros duplicados con los mismos valores en todos los campos
# (dado que no tenemos identificador) Y contamos cuántos son

nrow(heart_data[duplicated(heart_data), ])
```

```
## [1] 1
```

```
# Vemos los registros que están duplicados
heart_data[duplicated(heart_data), ]
```

```
##     age sex chest_pain_type resting_blood_pressure cholesterol
## 165  38   1               2                    138         175
##     fasting_blood_sugar rest_ecg_type max_heart_rate_achieved
## 165                   0            1                     173
##     exercise_induced_angina st_depression st_slope_type num_major_vessels
## 165                       0             0             2                 4
##     thalassemia_type heart_attack
## 165                2            1
```

Dado que solo existe un registro duplicado, con los mismos valores en todos los campos, no se va a eliminar porque es un porcentaje muy bajo del total y no afectará de manera significativa a los resultados que obtendremos más adelante. Además, al ser solo un registro, podría ser el caso de que esos dos pacientes fueran distintos y tuvieran las mismas características. Si tuviéramos muchos más, entonces seguramente serían los mismos pacientes y tendríamos que eliminarlos.

A continuación, se muestran algunos registros e información general de los datos que servirá para posteriomente proceder a la limpieza y conversión de los datos.

```
# Mostramos los tipos de datos de las variables tal y como las interpreta R
sapply(heart_data,class)
```

```
##                     age                     sex         chest_pain_type
##               "integer"               "integer"               "integer"
##  resting_blood_pressure             cholesterol     fasting_blood_sugar
##               "integer"               "integer"               "integer"
##           rest_ecg_type max_heart_rate_achieved exercise_induced_angina
##               "integer"               "integer"               "integer"
##           st_depression          st_slope_type       num_major_vessels
##               "numeric"               "integer"               "integer"
##        thalassemia_type            heart_attack
##               "integer"               "integer"
```

```
# Mostramos un resumen de los datos
summary(heart_data)
```

```
##       age             sex         chest_pain_type resting_blood_pressure
##  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
##   cholesterol    fasting_blood_sugar rest_ecg_type    max_heart_rate_achieved
##  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
##  exercise_induced_angina st_depression  st_slope_type   num_major_vessels
##  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
##  thalassemia_type  heart_attack
##  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
```

```
# Se muestran las 4 primeras observaciones de los datos
head(heart_data,4)
```

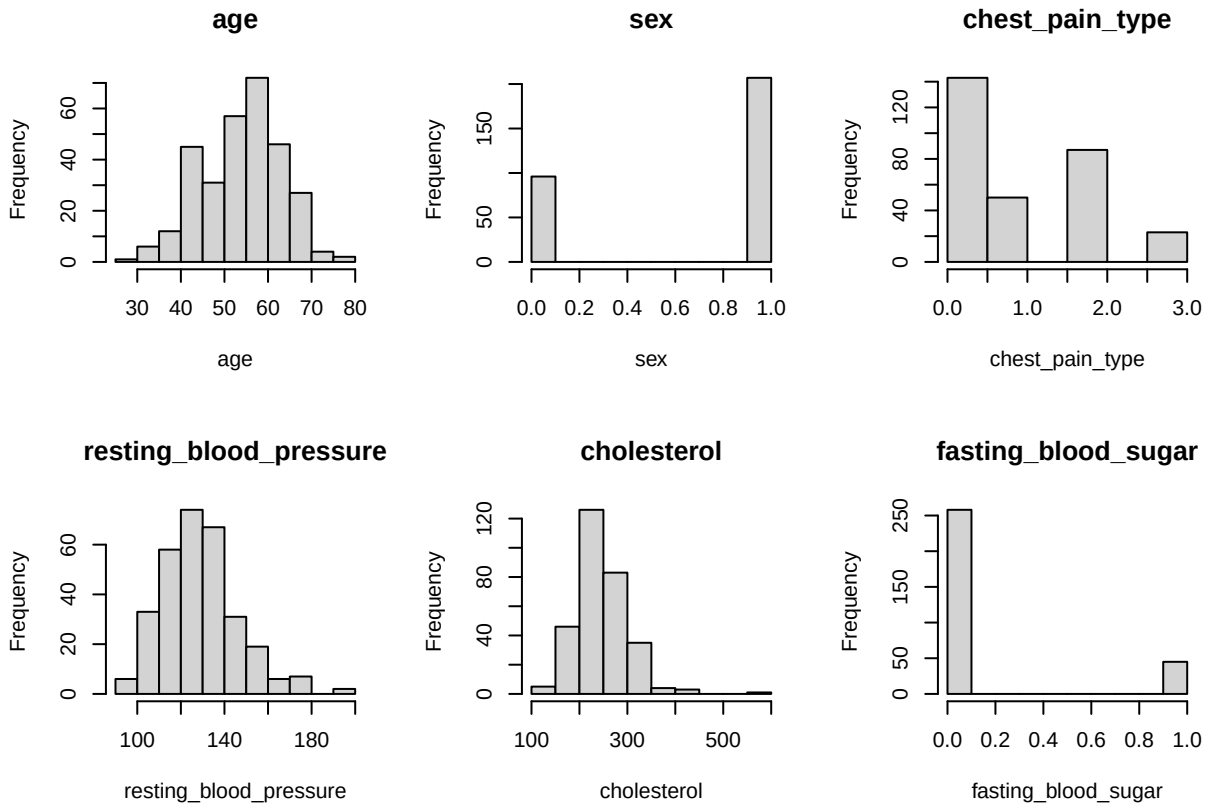
```
##   age sex chest_pain_type resting_blood_pressure cholesterol
## 1  63   1               3                    145         233
## 2  37   1               2                    130         250
## 3  41   0               1                    130         204
## 4  56   1               1                    120         236
##   fasting_blood_sugar rest_ecg_type max_heart_rate_achieved
## 1                   1             0                     150
## 2                   0             1                     187
## 3                   0             0                     172
## 4                   0             1                     178
##   exercise_induced_angina st_depression st_slope_type num_major_vessels
## 1                       0           2.3             0                 0
## 2                       0           3.5             0                 0
## 3                       0           1.4             2                 0
## 4                       0           0.8             2                 0
##   thalassemia_type heart_attack
## 1                1            1
## 2                2            1
## 3                2            1
## 4                2            1
```

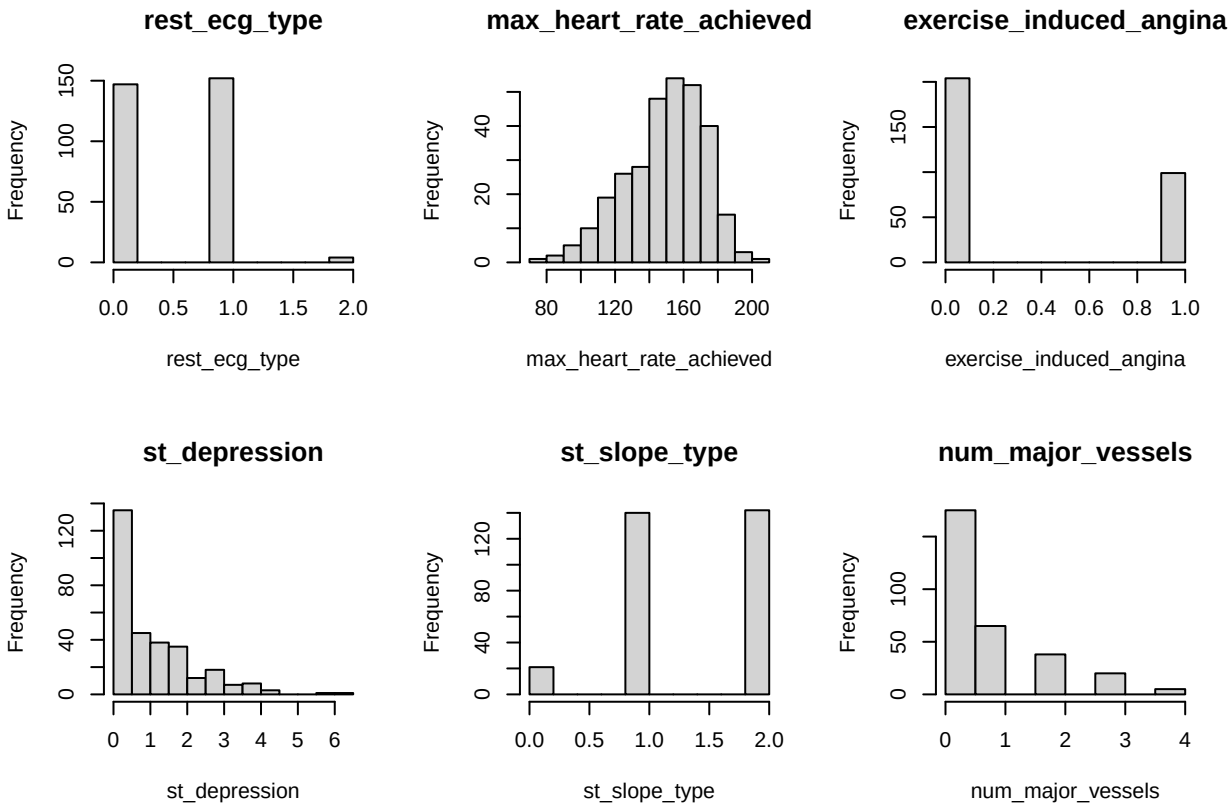
Por último, para nuestro análisis no se van a descartar registros porque no nos vamos a centrar en un tramo de edad concreto, sexo o una cantidad de colesterol, sino que se van a considerar a todos los pacientes con

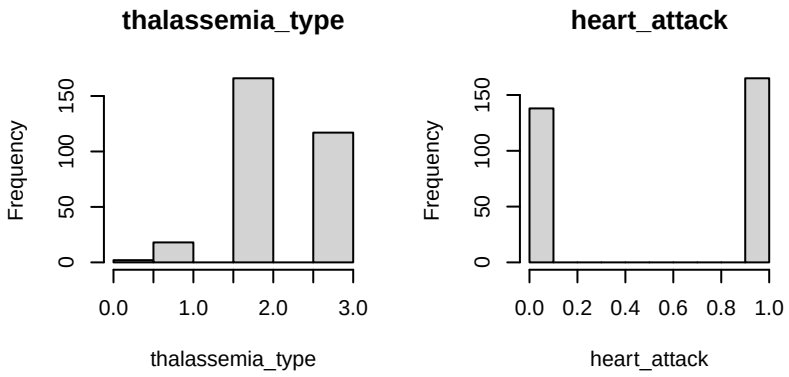
todas sus características para extraer el mayor número de conclusiones posibles teniendo en cuenta todos los atributos.

3 Visualización de los datos

```r
visualize_distribution <- function(variable) {
  # Seleccionamos la columna de la variable del conjunto de datos
  values <- heart_data[[variable]]
  # Creamos el histograma
  hist(values, xlab = variable, main = variable)
}
```




rest_ecg_type
Frequency
0
50
100
150
0.0
0.5
1.0
1.5
2.0
rest_ecg_type
max_heart_rate_achieved
Frequency
0
20
40
80
120
160
200
max_heart_rate_achieved
exercise_induced_angina
Frequency
0
50
150
0.0
0.2
0.4
0.6
0.8
1.0
exercise_induced_angina
st_depression
Frequency
0
40
80
120
0
1
2
3
4
5
6
st_depression
st_slope_type
Frequency
0
40
80
120
0.0
0.5
1.0
1.5
2.0
st_slope_type
num_major_vessels
Frequency
0
50
100
0
1
2
3
4
num_major_vessels



4 Limpieza de los datos.

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

4.1.1 Caso: Ceros

```
# Analisis de las columnas que contienen ceros en sus valores
cols_with_zeros <- which(apply(heart_data, 2, function(x) sum(x == 0)) > 0)
colnames(heart_data)[cols_with_zeros]
```

```
##  [1] "sex"                    "chest_pain_type"
##  [3] "fasting_blood_sugar"    "rest_ecg_type"
##  [5] "exercise_induced_angina" "st_depression"
##  [7] "st_slope_type"          "num_major_vessels"
##  [9] "thalassemia_type"       "heart_attack"
```

Las variables que contienen algún valor igual a cero son variables que esperan reflejar este valor tal y como se ha definido en el enunciado, por lo tanto no se va a realizar una limpieza de datos para este caso en particular. Véase a continuación las variables que aparecen con algún valor cero son las siguientes:

– "sex": Refleja el género de la persona (1 = masculino, 0 = femenino).

– “chest_pain_type”: Identifica el tipo de dolor en el pecho (0 = angina típica, 1 = angina atípica, 2 = dolor no anginoso, 3 = asintomático).

– “fasting_blood_sugar”: Indica si el nivel de azúcar en sangre en ayunas es mayor a 120 mg/dl (1 = verdadero, 0 = falso).

– “rest_ecg_type”: Muestra los resultados electrocardiográficos en reposo (0 = normal, 1 = anomalía de onda ST-T (inversiones de onda T y/o elevación o depresión ST de > 0,05 mV), 2 = hipertrofia ventricular izquierda probable o definida por los criterios de Estes).

– “exercise_induced_angina”: Indica si la angina ha sido inducida por el ejercicio (1 = sí, 0 = no).

– “st_depression”: Señala la depresión ST inducida por el ejercicio en relación con el descanso.

– “st_slope_type”: Muestra la pendiente del segmento ST de ejercicio máximo (0 = inclinación hacia abajo, 1 = plano, 2 = inclinación hacia arriba).

– “num_major_vessels”: Indica el número de buques principales (0, 1, 2, 3).

– “thalassemia_type”: Señala el ratio de un trastorno sanguíneo llamado talasemia (0 = no tiene, 1 = defecto fijo (sin flujo sanguíneo en alguna parte del corazón), 2 = flujo sanguíneo normal, 3 = defecto reversible (se observa un flujo sanguíneo, pero no es normal)).

– “heart_attack”: Indica si el paciente sufre un ataque al corazón o no (0 = No, 1 = Sí).

4.1.2 Caso: Elementos Vacíos

A continuación se realiza la comprobación de si hay elementos vacíos en el dataset, para cada columna se realiza el conteo de elementos vacíos existentes.

```
# Elementos vacíos de las variables del dataset
colSums(is.na(heart_data))
```

```
##                       age                       sex         chest_pain_type
##                         0                         0                       0
##  resting_blood_pressure               cholesterol     fasting_blood_sugar
##                         0                         0                       0
##           rest_ecg_type max_heart_rate_achieved exercise_induced_angina
##                         0                         0                       0
##           st_depression           st_slope_type       num_major_vessels
##                         0                         0                       0
##        thalassemia_type            heart_attack
##                         0                         0
```

Como se visualiza en los resultados anteriores no existen elementos vacíos en el conjunto de datos. Con ello, no será necesario realizar ningún procedimiento de limpieza de datos para valores vacíos de las variables del dataset.

4.1.3 Conversión y adaptación de los datos

Se van a realizar algunas conversiones de los tipos de algunas variables para realizar un análisis más eficiente y que nos facilite la interpretación de los resultados.

Primero convertiremos las siguientes variables numéricas a categóricas:

```r
# Transformamos a tipo factor las siguientes variables
heart_data$sex <- factor(heart_data$sex, levels = c(0,1), labels=
                           c("Femenino", "Masculino"))

heart_data$chest_pain_type <- factor(heart_data$chest_pain_type, levels = c(0,1,2,3), labels=
                                       c("Angina típica", "Angina atípica",
                                        "Dolor no anginoso","Asintomático"))

heart_data$fasting_blood_sugar <- factor(heart_data$fasting_blood_sugar, levels = c(0,1),
                                         labels=
                                           c("Azúcar Bajo", "Azúcar Alto"))

heart_data$rest_ecg_type <- factor(heart_data$rest_ecg_type, levels = c(0,1,2), labels=
                                     c("Normal", "Anomalía de onda ST-T",
                                      "Hipertrofia ventricular izquierda"))

heart_data$exercise_induced_angina <- factor(heart_data$exercise_induced_angina,
                                             levels = c(0,1), labels= c("No", "Sí"))

heart_data$st_slope_type <- factor(heart_data$st_slope_type, levels = c(0,1,2),
                                   labels= c("Baja", "Normal","Alta"))
heart_data$thalassemia_type <- factor(heart_data$thalassemia_type, levels = c(0,1,2,3),
                                      labels= c("Inexistente", "Fijo",
                                          "Normal","Reversible"))
```

También se pueden aplicar otro tipo de conversiones como por ejemplo la normalización *z-score* que resta la media a la variable y la divide por su desviación estándar. Usaremos esta normalización usando la función *scale* para normalizar las variables cuantitativas.

```r
# Indices de las variables cuantitativas
idx_var_cuant <- c(1,4,5,8,10,12)

# Normalización variables cuantitativas
heart_norm <- scale(heart_data[,idx_var_cuant])
```

Es posible que se tengan que utilizar más adelanete será estos datos normalizados, sin embargo, se van a mantener sin normalizar ya que para mostrar los resultados resulta más intuitivo verlos en su escala natural.

En el caso de las variables que no presenten una distribución normal, una opción sería realizar transformaciones de tipo Box-Cox para poder mejorar su normalidad y su homocedasticidad.

Asimismo, para algunas variables como por ejemplo la edad del paciente, sería interesante realizar un proceso de discretización. Esto nos permitiría agrupar las edades en diferentes grupos y poder sacar conclusiones que nos aporten un valor simbólico más allá de solo un número, aportándonos mayor información.

4.2 Identifica y gestiona los valores extremos

En primer lugar se realiza la visualización de los valores extremos para las variables: *"age", "cholesterol", "max_heart_rate_achieved", "resting_blood_pressure", "st_depression".*

```r
# Selección de las variables del conjunto de datos
data_selected <- subset(heart_data,
                select=c("age", "cholesterol", "max_heart_rate_achieved", "resting_blood_pressure", "st_
data_selected
```

```
##     age cholesterol max_heart_rate_achieved resting_blood_pressure
## 1    63         233                     150                    145
## 2    37         250                     187                    130
## 3    41         204                     172                    130
## 4    56         236                     178                    120
## 5    57         354                     163                    120
## 6    57         192                     148                    140
## 7    56         294                     153                    140
## 8    44         263                     173                    120
## 9    52         199                     162                    172
## 10   57         168                     174                    150
## 11   54         239                     160                    140
## 12   48         275                     139                    130
## 13   49         266                     171                    130
## 14   64         211                     144                    110
## 15   58         283                     162                    150
## 16   50         219                     158                    120
## 17   58         340                     172                    120
## 18   66         226                     114                    150
## 19   43         247                     171                    150
## 20   69         239                     151                    140
## 21   59         234                     161                    135
## 22   44         233                     179                    130
## 23   42         226                     178                    140
## 24   61         243                     137                    150
## 25   40         199                     178                    140
## 26   71         302                     162                    160
## 27   59         212                     157                    150
## 28   51         175                     123                    110
## 29   65         417                     157                    140
## 30   53         197                     152                    130
## 31   41         198                     168                    105
## 32   65         177                     140                    120
## 33   44         219                     188                    130
## 34   54         273                     152                    125
## 35   51         213                     125                    125
## 36   46         177                     160                    142
## 37   54         304                     170                    135
## 38   54         232                     165                    150
## 39   65         269                     148                    155
## 40   65         360                     151                    160
## 41   51         308                     142                    140
## 42   48         245                     180                    130
## 43   45         208                     148                    104
## 44   53         264                     143                    130
## 45   39         321                     182                    140
## 46   52         325                     172                    120
## 47   44         235                     180                    140
## 48   47         257                     156                    138
## 49   53         216                     115                    128
## 50   53         234                     160                    138
## 51   51         256                     149                    130
## 52   66         302                     151                    120
## 53   62         231                     146                    130
```

```
## 54   44         141              175              108
## 55   63         252              172              135
## 56   52         201              158              134
## 57   48         222              186              122
## 58   45         260              185              115
## 59   34         182              174              118
## 60   57         303              159              128
## 61   71         265              130              110
## 62   54         309              156              108
## 63   52         186              190              118
## 64   41         203              132              135
## 65   58         211              165              140
## 66   35         183              182              138
## 67   51         222              143              100
## 68   45         234              175              130
## 69   44         220              170              120
## 70   62         209              163              124
## 71   54         258              147              120
## 72   51         227              154               94
## 73   29         204              202              130
## 74   51         261              186              140
## 75   43         213              165              122
## 76   55         250              161              135
## 77   51         245              166              125
## 78   59         221              164              140
## 79   52         205              184              128
## 80   58         240              154              105
## 81   41         250              179              112
## 82   45         308              170              128
## 83   60         318              160              102
## 84   52         298              178              152
## 85   42         265              122              102
## 86   67         564              160              115
## 87   68         277              151              118
## 88   46         197              156              101
## 89   54         214              158              110
## 90   58         248              122              100
## 91   48         255              175              124
## 92   57         207              168              132
## 93   52         223              169              138
## 94   54         288              159              132
## 95   45         160              138              112
## 96   53         226              111              142
## 97   62         394              157              140
## 98   52         233              147              108
## 99   43         315              162              130
## 100  53         246              173              130
## 101  42         244              178              148
## 102  59         270              145              178
## 103  63         195              179              140
## 104  42         240              194              120
## 105  50         196              163              129
## 106  68         211              115              120
## 107  69         234              131              160
```

```
## 108  45         236                152                138
## 109  50         244                162                120
## 110  50         254                159                110
## 111  64         325                154                180
## 112  57         126                173                150
## 113  64         313                133                140
## 114  43         211                161                110
## 115  55         262                155                130
## 116  37         215                170                120
## 117  41         214                168                130
## 118  56         193                162                120
## 119  46         204                172                105
## 120  46         243                152                138
## 121  64         303                122                130
## 122  59         271                182                138
## 123  41         268                172                112
## 124  54         267                167                108
## 125  39         199                179                 94
## 126  34         210                192                118
## 127  47         204                143                112
## 128  67         277                172                152
## 129  52         196                169                136
## 130  74         269                121                120
## 131  54         201                163                160
## 132  49         271                162                134
## 133  42         295                162                120
## 134  41         235                153                110
## 135  41         306                163                126
## 136  49         269                163                130
## 137  60         178                 96                120
## 138  62         208                140                128
## 139  57         201                126                110
## 140  64         263                105                128
## 141  51         295                157                120
## 142  43         303                181                115
## 143  42         209                173                120
## 144  67         223                142                106
## 145  76         197                116                140
## 146  70         245                143                156
## 147  44         242                149                118
## 148  60         240                171                150
## 149  44         226                169                120
## 150  42         180                150                130
## 151  66         228                138                160
## 152  71         149                125                112
## 153  64         227                155                170
## 154  66         278                152                146
## 155  39         220                152                138
## 156  58         197                131                130
## 157  47         253                179                130
## 158  35         192                174                122
## 159  58         220                144                125
## 160  56         221                163                130
## 161  56         240                169                120
```

```
## 162  55         342                166                132
## 163  41         157                182                120
## 164  38         175                173                138
## 165  38         175                173                138
## 166  67         286                108                160
## 167  67         229                129                120
## 168  62         268                160                140
## 169  63         254                147                130
## 170  53         203                155                140
## 171  56         256                142                130
## 172  48         229                168                110
## 173  58         284                160                120
## 174  58         224                173                132
## 175  60         206                132                130
## 176  40         167                114                110
## 177  60         230                160                117
## 178  64         335                158                140
## 179  43         177                120                120
## 180  57         276                112                150
## 181  55         353                132                132
## 182  65         225                114                150
## 183  61         330                169                130
## 184  58         230                165                112
## 185  50         243                128                150
## 186  44         290                153                112
## 187  60         253                144                130
## 188  54         266                109                124
## 189  50         233                163                140
## 190  41         172                158                110
## 191  51         305                142                130
## 192  58         216                131                128
## 193  54         188                113                120
## 194  60         282                142                145
## 195  60         185                155                140
## 196  59         326                140                170
## 197  46         231                147                150
## 198  67         254                163                125
## 199  62         267                 99                120
## 200  65         248                158                110
## 201  44         197                177                110
## 202  60         258                141                125
## 203  58         270                111                150
## 204  68         274                150                180
## 205  62         164                145                160
## 206  52         255                161                128
## 207  59         239                142                110
## 208  60         258                157                150
## 209  49         188                139                120
## 210  59         177                162                140
## 211  57         229                150                128
## 212  61         260                140                120
## 213  39         219                140                118
## 214  61         307                146                145
## 215  56         249                144                125
```

```
## 216  43         341              136              132
## 217  62         263               97              130
## 218  63         330              132              130
## 219  65         254              127              135
## 220  48         256              150              130
## 221  63         407              154              150
## 222  55         217              111              140
## 223  65         282              174              138
## 224  56         288              133              200
## 225  54         239              126              110
## 226  70         174              125              145
## 227  62         281              103              120
## 228  35         198              130              120
## 229  59         288              159              170
## 230  64         309              131              125
## 231  47         243              152              108
## 232  57         289              124              165
## 233  55         289              145              160
## 234  64         246               96              120
## 235  70         322              109              130
## 236  51         299              173              140
## 237  58         300              171              125
## 238  60         293              170              140
## 239  77         304              162              125
## 240  35         282              156              126
## 241  70         269              112              160
## 242  59         249              143              174
## 243  64         212              132              145
## 244  57         274               88              152
## 245  56         184              105              132
## 246  48         274              166              124
## 247  56         409              150              134
## 248  66         246              120              160
## 249  54         283              195              192
## 250  69         254              146              140
## 251  51         298              122              140
## 252  43         247              143              132
## 253  62         294              106              138
## 254  67         299              125              100
## 255  59         273              125              160
## 256  45         309              147              142
## 257  58         259              130              128
## 258  50         200              126              144
## 259  62         244              154              150
## 260  38         231              182              120
## 261  66         228              165              178
## 262  52         230              160              112
## 263  53         282               95              123
## 264  63         269              169              108
## 265  54         206              108              110
## 266  66         212              132              112
## 267  55         327              117              180
## 268  49         149              126              118
## 269  54         286              116              122
```

```
## 270  56         283                  103                  130
## 271  46         249                  144                  120
## 272  61         234                  145                  134
## 273  67         237                   71                  120
## 274  58         234                  156                  100
## 275  47         275                  118                  110
## 276  52         212                  168                  125
## 277  58         218                  105                  146
## 278  57         261                  141                  124
## 279  58         319                  152                  136
## 280  61         166                  125                  138
## 281  42         315                  125                  136
## 282  52         204                  156                  128
## 283  59         218                  134                  126
## 284  40         223                  181                  152
## 285  61         207                  138                  140
## 286  46         311                  120                  140
## 287  59         204                  162                  134
## 288  57         232                  164                  154
## 289  57         335                  143                  110
## 290  55         205                  130                  128
## 291  61         203                  161                  148
## 292  58         318                  140                  114
## 293  58         225                  146                  170
## 294  67         212                  150                  152
## 295  44         169                  144                  120
## 296  63         187                  144                  140
## 297  63         197                  136                  124
## 298  59         176                   90                  164
## 299  57         241                  123                  140
## 300  45         264                  132                  110
## 301  68         193                  141                  144
## 302  57         131                  115                  130
## 303  57         236                  174                  130
##     st_depression
## 1             2.3
## 2             3.5
## 3             1.4
## 4             0.8
## 5             0.6
## 6             0.4
## 7             1.3
## 8             0.0
## 9             0.5
## 10            1.6
## 11            1.2
## 12            0.2
## 13            0.6
## 14            1.8
## 15            1.0
## 16            1.6
## 17            0.0
## 18            2.6
## 19            1.5
```

```
## 20           1.8
## 21           0.5
## 22           0.4
## 23           0.0
## 24           1.0
## 25           1.4
## 26           0.4
## 27           1.6
## 28           0.6
## 29           0.8
## 30           1.2
## 31           0.0
## 32           0.4
## 33           0.0
## 34           0.5
## 35           1.4
## 36           1.4
## 37           0.0
## 38           1.6
## 39           0.8
## 40           0.8
## 41           1.5
## 42           0.2
## 43           3.0
## 44           0.4
## 45           0.0
## 46           0.2
## 47           0.0
## 48           0.0
## 49           0.0
## 50           0.0
## 51           0.5
## 52           0.4
## 53           1.8
## 54           0.6
## 55           0.0
## 56           0.8
## 57           0.0
## 58           0.0
## 59           0.0
## 60           0.0
## 61           0.0
## 62           0.0
## 63           0.0
## 64           0.0
## 65           0.0
## 66           1.4
## 67           1.2
## 68           0.6
## 69           0.0
## 70           0.0
## 71           0.4
## 72           0.0
## 73           0.0
```

```
## 74           0.0
## 75           0.2
## 76           1.4
## 77           2.4
## 78           0.0
## 79           0.0
## 80           0.6
## 81           0.0
## 82           0.0
## 83           0.0
## 84           1.2
## 85           0.6
## 86           1.6
## 87           1.0
## 88           0.0
## 89           1.6
## 90           1.0
## 91           0.0
## 92           0.0
## 93           0.0
## 94           0.0
## 95           0.0
## 96           0.0
## 97           1.2
## 98           0.1
## 99           1.9
## 100          0.0
## 101          0.8
## 102          4.2
## 103          0.0
## 104          0.8
## 105          0.0
## 106          1.5
## 107          0.1
## 108          0.2
## 109          1.1
## 110          0.0
## 111          0.0
## 112          0.2
## 113          0.2
## 114          0.0
## 115          0.0
## 116          0.0
## 117          2.0
## 118          1.9
## 119          0.0
## 120          0.0
## 121          2.0
## 122          0.0
## 123          0.0
## 124          0.0
## 125          0.0
## 126          0.7
## 127          0.1
```

```
## 128          0.0
## 129          0.1
## 130          0.2
## 131          0.0
## 132          0.0
## 133          0.0
## 134          0.0
## 135          0.0
## 136          0.0
## 137          0.0
## 138          0.0
## 139          1.5
## 140          0.2
## 141          0.6
## 142          1.2
## 143          0.0
## 144          0.3
## 145          1.1
## 146          0.0
## 147          0.3
## 148          0.9
## 149          0.0
## 150          0.0
## 151          2.3
## 152          1.6
## 153          0.6
## 154          0.0
## 155          0.0
## 156          0.6
## 157          0.0
## 158          0.0
## 159          0.4
## 160          0.0
## 161          0.0
## 162          1.2
## 163          0.0
## 164          0.0
## 165          0.0
## 166          1.5
## 167          2.6
## 168          3.6
## 169          1.4
## 170          3.1
## 171          0.6
## 172          1.0
## 173          1.8
## 174          3.2
## 175          2.4
## 176          2.0
## 177          1.4
## 178          0.0
## 179          2.5
## 180          0.6
## 181          1.2
```

```
## 182          1.0
## 183          0.0
## 184          2.5
## 185          2.6
## 186          0.0
## 187          1.4
## 188          2.2
## 189          0.6
## 190          0.0
## 191          1.2
## 192          2.2
## 193          1.4
## 194          2.8
## 195          3.0
## 196          3.4
## 197          3.6
## 198          0.2
## 199          1.8
## 200          0.6
## 201          0.0
## 202          2.8
## 203          0.8
## 204          1.6
## 205          6.2
## 206          0.0
## 207          1.2
## 208          2.6
## 209          2.0
## 210          0.0
## 211          0.4
## 212          3.6
## 213          1.2
## 214          1.0
## 215          1.2
## 216          3.0
## 217          1.2
## 218          1.8
## 219          2.8
## 220          0.0
## 221          4.0
## 222          5.6
## 223          1.4
## 224          4.0
## 225          2.8
## 226          2.6
## 227          1.4
## 228          1.6
## 229          0.2
## 230          1.8
## 231          0.0
## 232          1.0
## 233          0.8
## 234          2.2
## 235          2.4
```

```
## 236          1.6
## 237          0.0
## 238          1.2
## 239          0.0
## 240          0.0
## 241          2.9
## 242          0.0
## 243          2.0
## 244          1.2
## 245          2.1
## 246          0.5
## 247          1.9
## 248          0.0
## 249          0.0
## 250          2.0
## 251          4.2
## 252          0.1
## 253          1.9
## 254          0.9
## 255          0.0
## 256          0.0
## 257          3.0
## 258          0.9
## 259          1.4
## 260          3.8
## 261          1.0
## 262          0.0
## 263          2.0
## 264          1.8
## 265          0.0
## 266          0.1
## 267          3.4
## 268          0.8
## 269          3.2
## 270          1.6
## 271          0.8
## 272          2.6
## 273          1.0
## 274          0.1
## 275          1.0
## 276          1.0
## 277          2.0
## 278          0.3
## 279          0.0
## 280          3.6
## 281          1.8
## 282          1.0
## 283          2.2
## 284          0.0
## 285          1.9
## 286          1.8
## 287          0.8
## 288          0.0
## 289          3.0
```

```
## 290          2.0
## 291          0.0
## 292          4.4
## 293          2.8
## 294          0.8
## 295          2.8
## 296          4.0
## 297          0.0
## 298          1.0
## 299          0.2
## 300          1.2
## 301          3.4
## 302          1.2
## 303          0.0
```

```
# Función para simplificar la extracción de información de valores extremos
outlier_info <- function(var, name_var) {
  # Valores extremos en formato boxplot de la variable
  box = boxplot(var, main = name_var,
        ylab="Valor", col = "lightblue", horizontal = FALSE, outline = TRUE)
  # Identificar los valores atípicos
  outliers <- boxplot.stats(var)$out
  if (length(outliers) == 0) {
    cat("No se han identificado valores atípicos", "\n")
  } else {
    # Imprimir el número de valores atípicos
    cat("Outliers identificados:", unique(outliers), "\n")
  }
  # Imprimir los valores máximo y mínimo de los valores atípicos
  cat("Máximos y mínimos de los valores extremos identificados:", range(var))
}
```

```
par(mfrow=c(2, 3))
outlier_info(data_selected$age, "age")
```

```
## No se han identificado valores atípicos
## Máximos y mínimos de los valores extremos identificados: 29 77
```

```
outlier_info(data_selected$cholesterol, "cholesterol")
```

```
## Outliers identificados: 417 564 394 407 409
## Máximos y mínimos de los valores extremos identificados: 126 564
```

```
outlier_info(data_selected$max_heart_rate_achieved, "max_heart_rate_achieved")
```

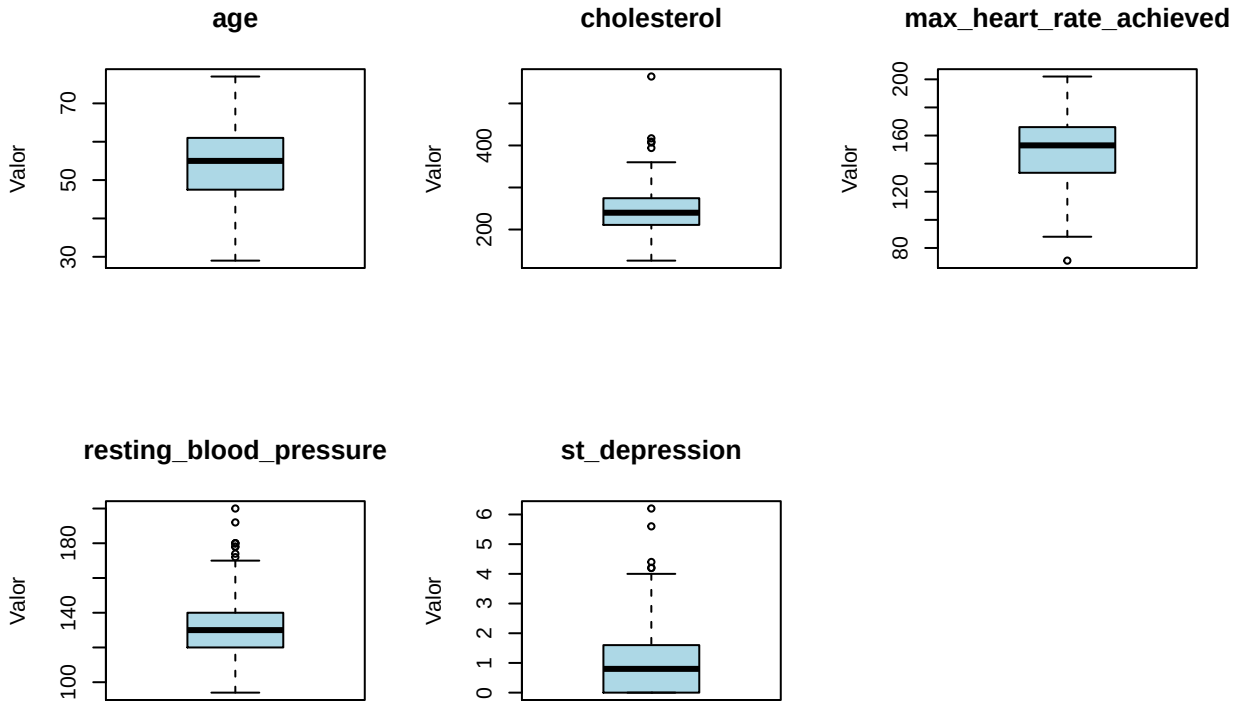
```
## Outliers identificados: 71
## Máximos y mínimos de los valores extremos identificados: 71 202
```

```
outlier_info(data_selected$resting_blood_pressure, "resting_blood_pressure")
```

```
## Outliers identificados: 172 178 180 200 174 192
## Máximos y mínimos de los valores extremos identificados: 94 200
```

```
outlier_info(data_selected$st_depression, "st_depression")
```

```
## Outliers identificados: 4.2 6.2 5.6 4.4
## Máximos y mínimos de los valores extremos identificados: 0 6.2
```



A continuación vamos a extraer las conclusiones pertinentes respectos a los valores extremos detectados en los resultados y los gráficos previos:

- En la variable "age", no se han identificado valores atípicos. Los valores máximo y mínimo de la variable son 29 y 77, respectivamente.
- En la variable "cholesterol", se han identificado 5 valores atípicos (outliers). Los valores máximo y mínimo de los outliers identificados son 126 y 564, respectivamente.
- En la variable "max_heart_rate_achieved", se ha identificado 1 valor atípico (outlier). Los valores máximo y mínimo de los outliers identificados son 71 y 202, respectivamente.
- En la variable "resting_blood_pressure", se han identificado 6 valores atípicos (outliers). Los valores máximo y mínimo de los outliers identificados son 94 y 200, respectivamente.
- En la variable "st_depression", se han identificado 4 valores atípicos (outliers). Los valores máximo y mínimo de los outliers identificados son 0 y 6.2, respectivamente.

Estos resultados indican que algunas de las variables tienen valores extremos que se alejan significativamente del resto y que pueden afectar el rendimiento de algunos algoritmos de análisis de datos.

5 Análisis de los datos.

5.1 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?).

==== JOAN ====

En este caso, se quiere analizar el conjunto de datos "heart.csv" para predecir si un paciente tiene un ataque al corazón o no. El conjunto de datos "o2Saturation.csv" no parece estar relacionado con este propósito y, por tanto, no se incluiría en el análisis.

Para analizar el conjunto de datos "heart.csv", se podría utilizar un modelo de aprendizaje supervisado para entrenar un modelo que use las variables independientes (edad, género, tipo de dolor en el pecho, etc.) como entrada y la variable dependiente (ataque al corazón o no) como salida. Una vez entrenado el modelo, se puede evaluar su desempeño y utilizarlo para hacer predicciones sobre pacientes futuros.

También se pueden comparar los resultados de distintos modelos de aprendizaje supervisado para ver cuál tiene mejor desempeño en este problema en particular. También se puede comparar el desempeño del modelo entrenado con diferentes subconjuntos de datos (por ejemplo, separando los pacientes por género o por edad).

Se podría utilizar un árbol de decisión para construir un modelo que use las variables independientes (edad, género, tipo de dolor en el pecho, etc.) como entrada y la variable dependiente (ataque al corazón o no) como salida. Una vez entrenado el modelo, se podría utilizar para hacer predicciones sobre pacientes futuros.

Para evaluar el desempeño del modelo, se podrían utilizar métricas como la precisión, la sensibilidad o el valor F1. También se podría comparar el desempeño del árbol de decisión con otros modelos de aprendizaje supervisado para ver cuál tiene mejor desempeño en este problema en particular.

Además, podríamos utilizar el test Chi cuadrado para ver si existe alguna asociación entre el género de un paciente y el resultado (ataque al corazón o no).

Para utilizar el test Chi cuadrado, necesitaríamos crear una tabla de contingencia con las frecuencias absolutas o relativas de cada combinación de variables. Luego, se calcularía el valor Chi cuadrado y se compararía con una tabla de valores críticos para determinar si existe una asociación significativa entre las variables.

5.2 Comprobación de la normalidad y homogeneidad de la varianza.

!#TODO: ESTO ES UN EJEMPLO DE PARTIDA

Para comprobar la normalidad de los datos, utilizamos la función `shapiro.test()`. Esta función toma un vector de datos y realiza un test de normalidad de Shapiro-Wilk. Si el p-valor devuelto es superior al nivel de significación, entonces se puede concluir que los datos siguen una distribución normal.

```
# Carga el conjunto de datos
data(iris)

# Extrae la longitud del conjunto de datos y realiza un test de normalidad
shapiro.test(iris$Sepal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Sepal.Length
## W = 0.97609, p-value = 0.01018
```

Para comprobar la homogeneidad de la varianza, utilizamos la función `var.test()`. Esta función toma dos vectores de datos y realiza un test de igualdad de varianzas. Si el p-valor devuelto es superior al nivel de significación que hayas elegido, entonces se puede concluir que las varianzas de los dos conjuntos de datos son iguales.

```
# Carga el conjunto de datos
data(iris)

# Compara la varianza
var.test(iris$Sepal.Length, iris$Petal.Length)
```

```
##
##  F test to compare two variances
##
## data:  iris$Sepal.Length and iris$Petal.Length
## F = 0.22004, num df = 149, denom df = 149, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1594015 0.3037352
## sample estimates:
## ratio of variances
##          0.2200361
```

5.3 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.

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

Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

7 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

8 Código.

Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

9 Vídeo.

Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace al Google Drive de la UOC, junto con enlace al repositorio Git entregafo.