A1: Preproceso de datos

Estadística Avanzada, Universitat Oberta de Catalunya

Paula Muñoz Lago

21 octubre 2019

1. Carga del fichero

El fichero de datos llamado "DataEstradiol.csv" se encuentra en la misma carpeta que el script de R, que a su vez ha sido previamente escogida como directorio de trabajo. Por lo que con los siguientes comandos obtendremos una variable de tipo data data frame, que guardaremos en la variable *data*. Dado que el formato del csv que queremos cargar es el español, es decir, separado por puntos y comas, utilizaremos la función *read.csv2*.

```
current_working_directory <- getwd()
data <- read.csv2(paste(current_working_directory,"/DataEstradiol.csv", sep = ""))
```

Para explorar el *data frame* recién cargado, utilizaremos el siguiente código.

```
nrow(data) #Numero de registros
```

```
## [1] 211
```

```
ncol(data) #Numero de variables
```

```
## [1] 11
```

```
colnames(data) #Nombre de las variables
```

```
##  [1] "Id"       "Estradl"  "Ethnic"   "Entage"   "Numchild" "Agefbo"
##  [7] "Anykids"  "Agemenar" "BMI"      "WHR"      "Area"
```

2. Breve inspección de los datos

Como resumen del apartado anterior, con la siguiente línea de código obtendremos toda la información previamente obtenida, adenñas del tipo de datos de cada variable y los primeros valores de cada una.

```
str(data)
```

```
## 'data.frame':    211 obs. of  11 variables:
##  $ Id      : int  2 2 3 6 8 9 11 13 14 15 ...
##  $ Estradl : Factor w/ 197 levels "10.1","10.7",..: 194 184 35 115 117 144 83 31 164 174 ...
##  $ Ethnic  : Factor w/ 8 levels "   African American   ",..: 7 7 2 7 2 6 7 2 2 2 ...
##  $ Entage  : Factor w/ 25 levels "11.0","12.0",..: 18 11 9 21 19 24 20 17 12 23 ...
##  $ Numchild: int  0 0 0 0 0 2 9 4 0 0 ...
##  $ Agefbo  : int  0 0 0 0 0 27 99 19 0 8 ...
##  $ Anykids : int  0 0 0 0 0 1 9 1 0 0 ...
##  $ Agemenar: Factor w/ 26 levels "10,0","10.0",..: 4 16 11 14 11 5 11 12 11 10 ...
##  $ BMI     : num  18.9 20.4 22.3 20.5 24.3 ...
##  $ WHR     : num  0.7 0.7 0.75 0.73 0.75 0.71 0.73 0.69 0.68 0.71 ...
##  $ Area    : int  0 0 1 1 1 1 1 1 1 1 ...
```

Sin embargo, si quisiese verse únicamente el tipo de las variables, podríamos hacerlo de la siguiente manera.

```
sapply(data, class)
```

```
##        Id   Estradl    Ethnic    Entage  Numchild    Agefbo   Anykids
## "integer"  "factor"  "factor"  "factor" "integer" "integer" "integer"
##  Agemenar       BMI       WHR      Area
##  "factor" "numeric" "numeric" "integer"
```

3. Formato de las variables cuantitativas

Explorando los datos en el apartado anterior, descubrimos que las variables cuantitativas son todas menos la que representa la Étnia (columna número 3). De esta forma creamos un array con el índice de dichas columnas y, en primer lugar, corregimos las posibles inconsistencias en el punto decimal, cambiando las comas que encontremos por puntos gracias a la función gsub.

```
variables_cuantitativas <- c(1, 2, 4:ncol(data))
data[, variables_cuantitativas] <- sapply(data[, variables_cuantitativas], gsub, pattern=",", replaceme
```

Una vez homogeneizado el sistema de representación decimal a la separación con punto, convertimos a tipo numérico todas las variables cuantitativas.

```
data[, variables_cuantitativas] <- sapply(data[, variables_cuantitativas], as.numeric)
sapply(data, class)
```

```
##        Id   Estradl    Ethnic    Entage  Numchild    Agefbo   Anykids
## "numeric" "numeric"  "factor" "numeric" "numeric" "numeric" "numeric"
##  Agemenar       BMI       WHR      Area
## "numeric" "numeric" "numeric" "numeric"
```

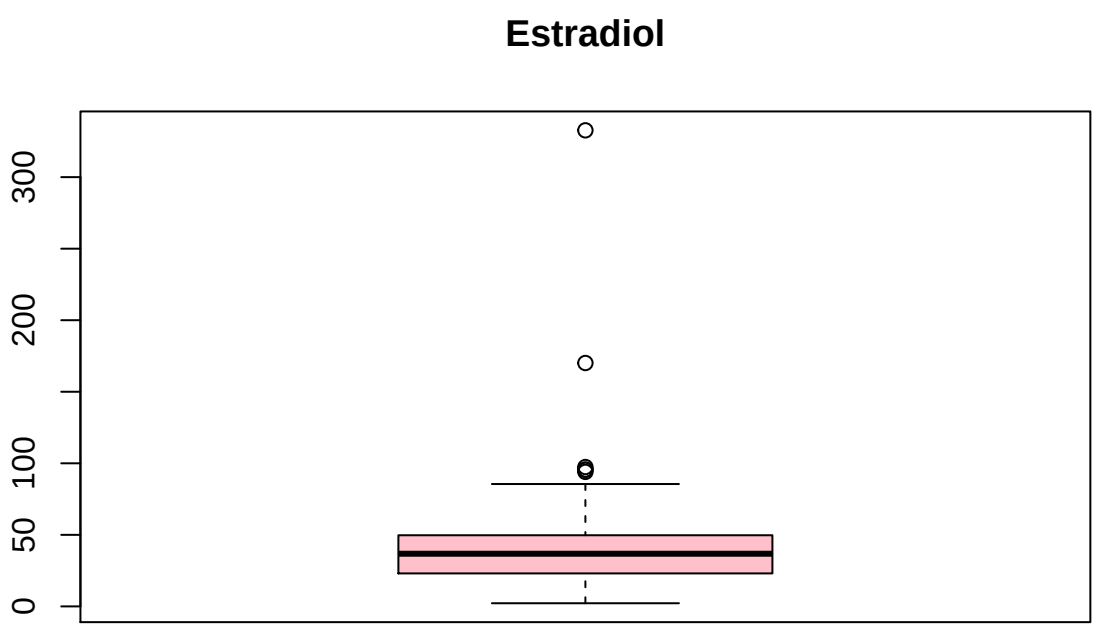
Antes de continuar, será de gran utilidad normalizar el valor centinela de cada variable, por ejemplo, en *Anykids* el valor centinela es 9, mientras que en *NumChild* es 99. Asignaremos a cada uno el valor *NA* (Not available), e imputaremos dichos valores perdidos en el punto 8

```
data$Numchild[which(data$Numchild == 9)] <- NA
data$Agefbo[which(data$Agefbo == 99)] <- NA
data$Anykids[which(data$Anykids == 9)] <- NA
data$Agemenar[which(data$Agemenar == 99)] <- NA
```

4. Valores extremos

Para detectar la presencia de valores extremos en la variable *Estradiol*, imprimimos en primer lugar la siguiente gráfica de caja para visualizar la distribución de los valores.

```
boxplot(data$Estradl,main="Estradiol", col="pink")
```

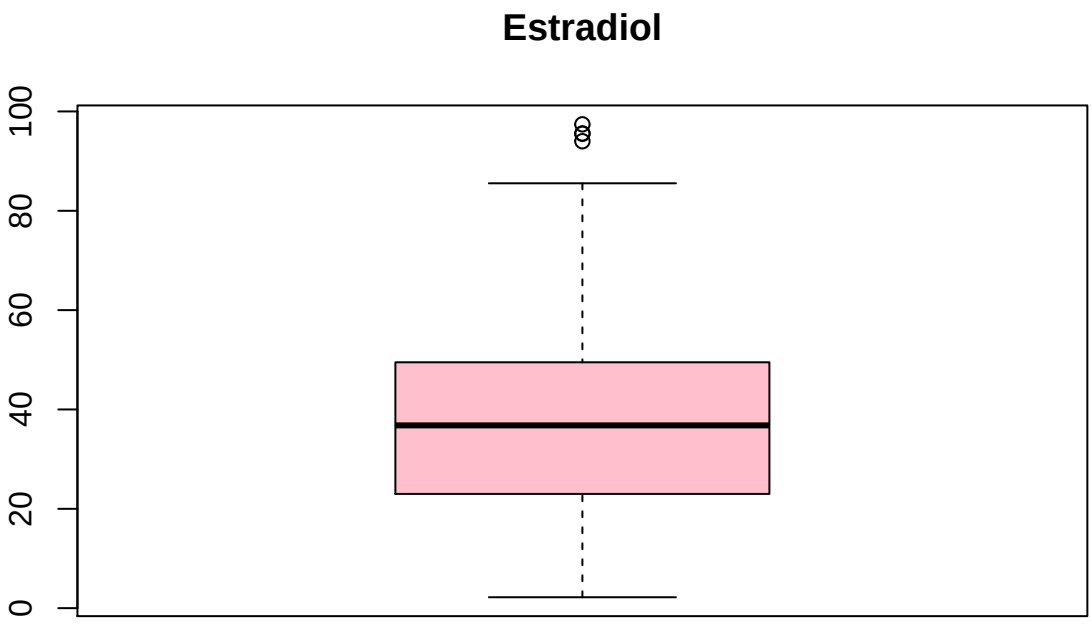


Podemos ver que el rango intercuartílico engloba valores desde el 25 hasta algo más de 50. Como se aprecia en la representación, existen algunos valores atípicos. Podemos descubrir cuales son de la siguiente manera.

```
outliers <- sort(boxplot.stats(data$Estradl)$out)
```

Para tomar la decisión de si tenemos que eliminar los registros correspondientes a algunos de los valores atípicos previamente obtenidos, podemos, en primer lugar, informarmos de cuales son los valores que toma el Estradiol normalmente [https://www.reproduccionasistida.org/valores-hormonales-en-la-mujer/][1]. Puesto que la información plasmada en la página web indica que los valores del estradiol van desde los 27 pg/ml hasta los 161 pg/ml, consideramos que los valores 170.1 y 332.70 son erroneos. De esta forma, extraeremos en primer lugar el índice de dichos registros para a continuación eliminarlos del *data frame*.

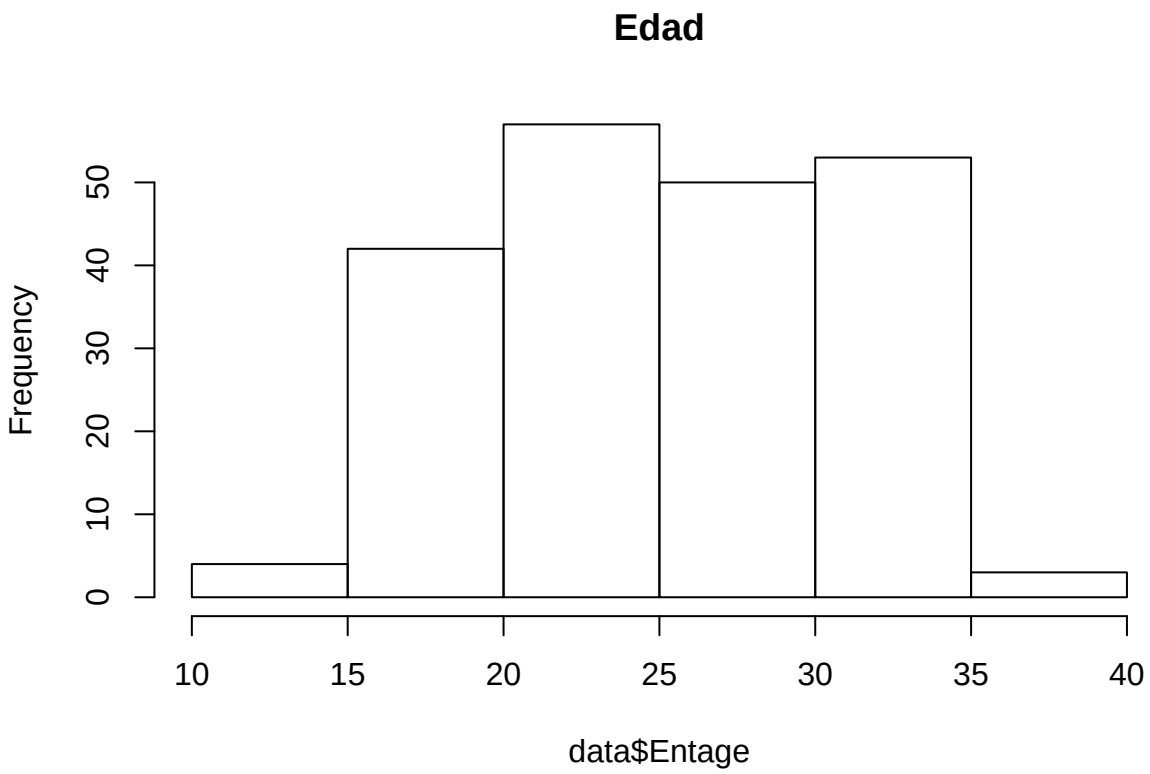
```
remove_index <- which(data$Estradl %in% tail(outliers, 2))
data <- data[-remove_index,]
boxplot(data$Estradl,main="Estradiol", col="pink")
```



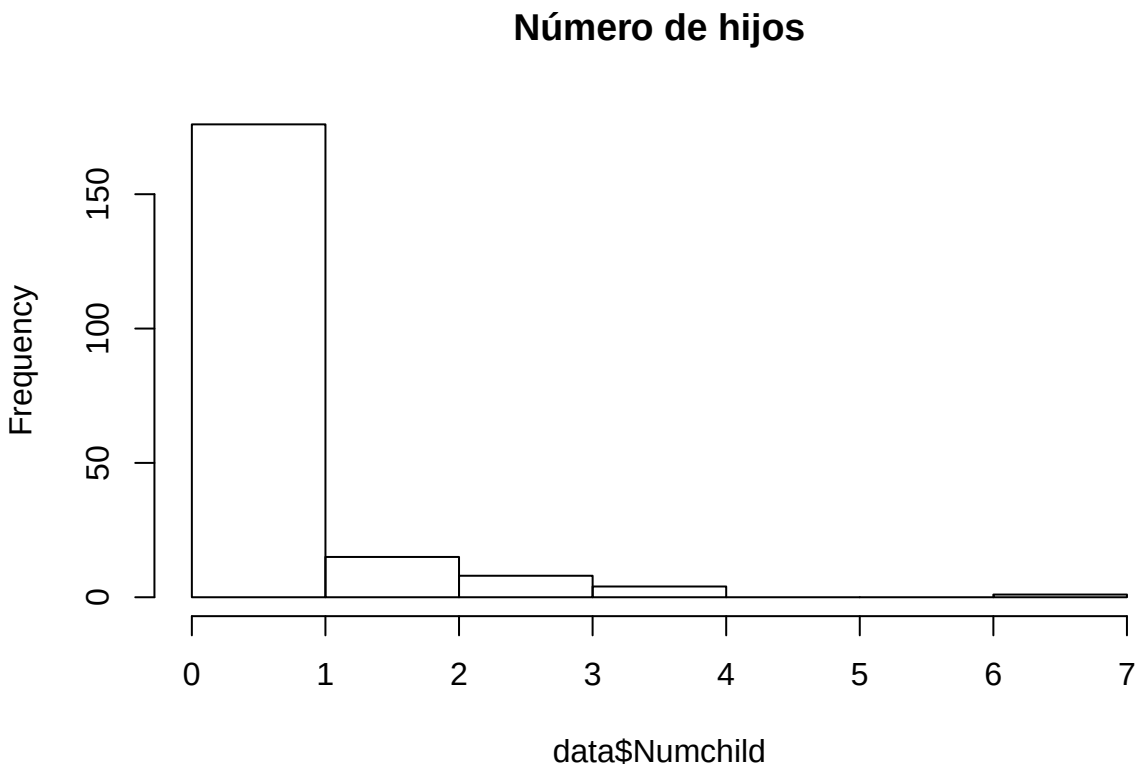
5. Valores del resto de variables cuantitativas

Procedemos a inspeccionar el resto de variables cuantitativas en busca de algún valor anómalo

```
hist(data$Entage,main="Edad")
```



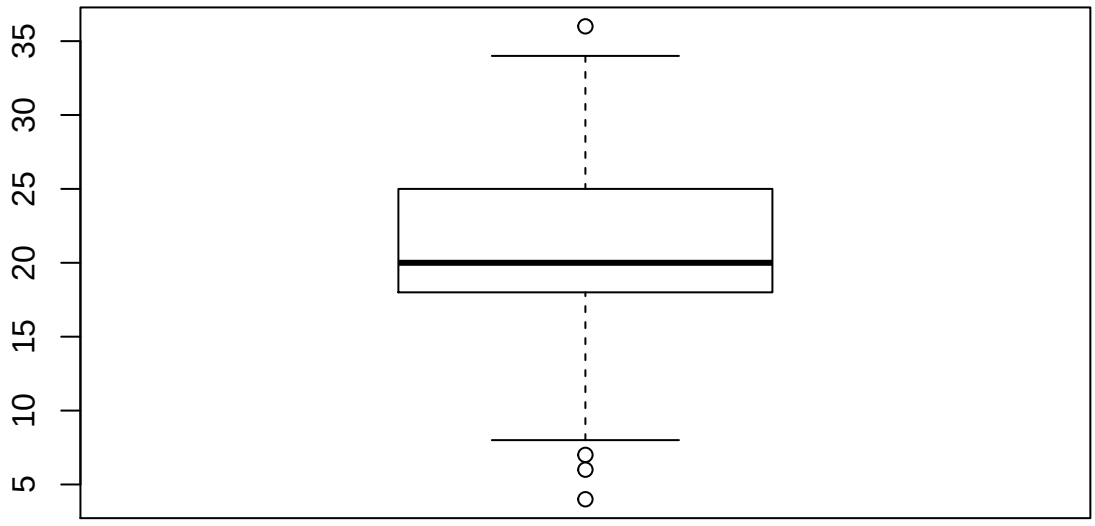

```
hist(data$Numchild,main="Número de hijos")
```



La edad consta desde 10 hasta 40 años, mientras que el número de hijos desde 0 hasta 7. En estos casos, los valores atípicos pueden no estar dados por un fallo al introducir los datos, por lo que mantendremos los datos como están.

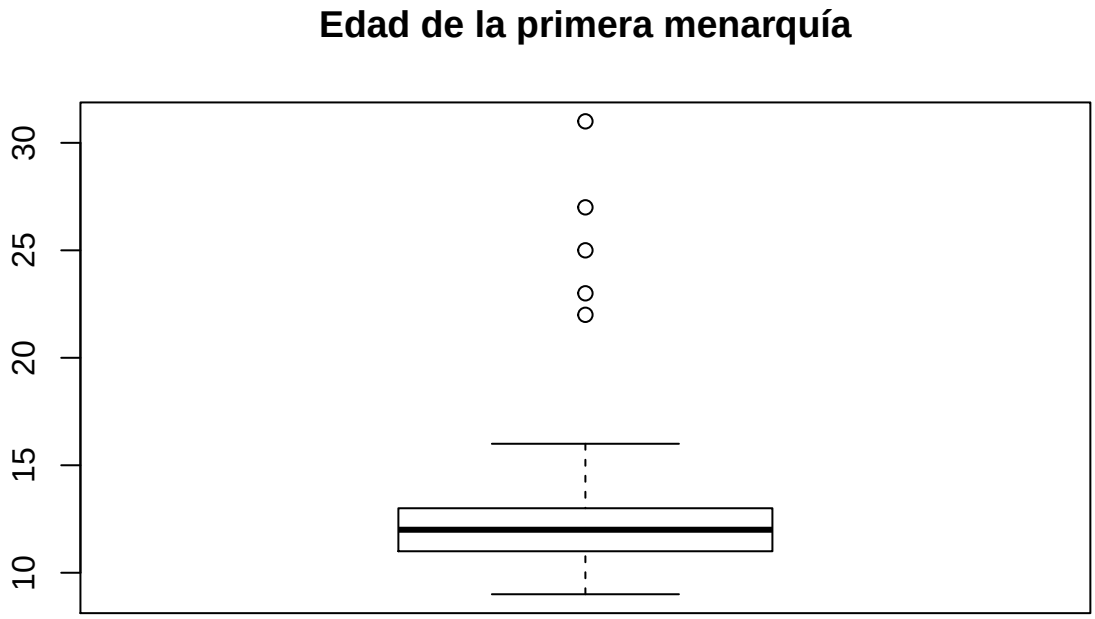
En el caso de la edad del primer hijo, la mayoría de entradas contienen el dato "0", lo cual indica que no han tenido hijos, y evitaremos visualizar ese dato. Podemos ver que en algunos registros aparece que el primer hijo se tuvo a los 4, 6, y 7 años. Puesto que se trata de datos erroneos, ya que no es posible, asignaremos el valor *NA* y los imputaremos en el punto 8.

```
boxplot(data$Agefbo[which(data$Agefbo != 0)],main="Edad del primer hijo")
anomalias <- boxplot(data$Agefbo[which(data$Agefbo != 0)],main="Edad del primer hijo")$out
```




```
data$Agefbo[which(data$Agefbo == 4 | data$Agefbo == 6 | data$Agefbo == 7)] <- NA
```

```
boxplot(data$Agemenar,main="Edad de la primera menarquía")
boxplot(data$Agemenar,main="Edad de la primera menarquía")$out
```

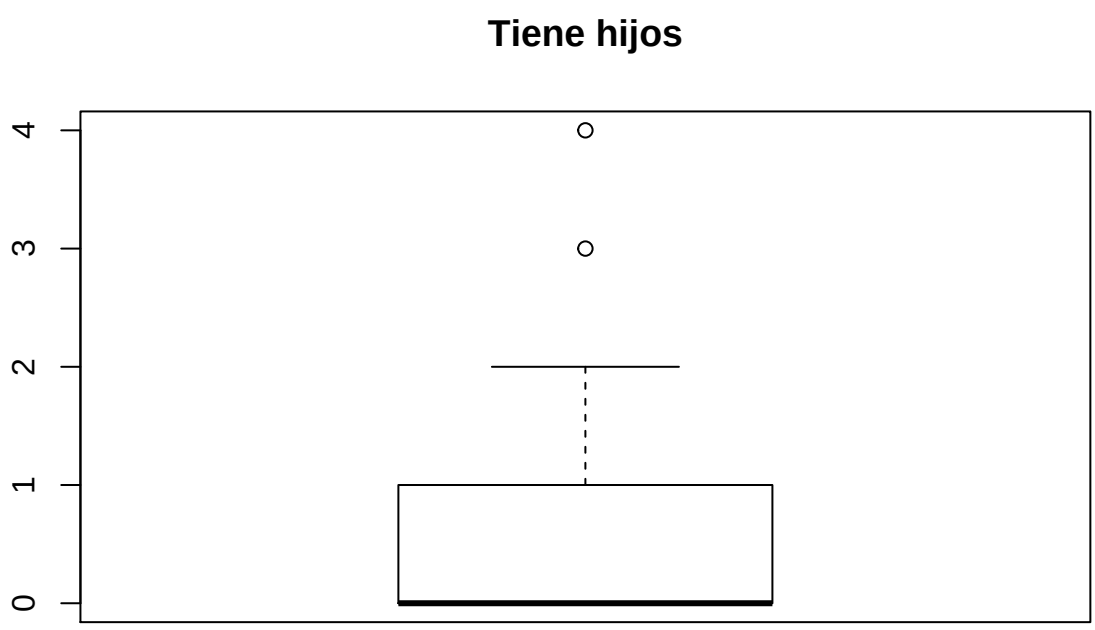



```
## [1] 31 22 25 23 27
```

Los valores anómalos los corregiremos en el punto 6.

En el caso de la variable booleana *Anykids*, representa si la mujer ha tenido hijos o no, y todos sus valores se distribuyen entre el valor 0, que indica que no ha tenido hijos, y el 1, que indica lo contrario.

```
boxplot(data$Anykids,main="Tiene hijos")
boxplot(data$Anykids,main="Tiene hijos")$out
```




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

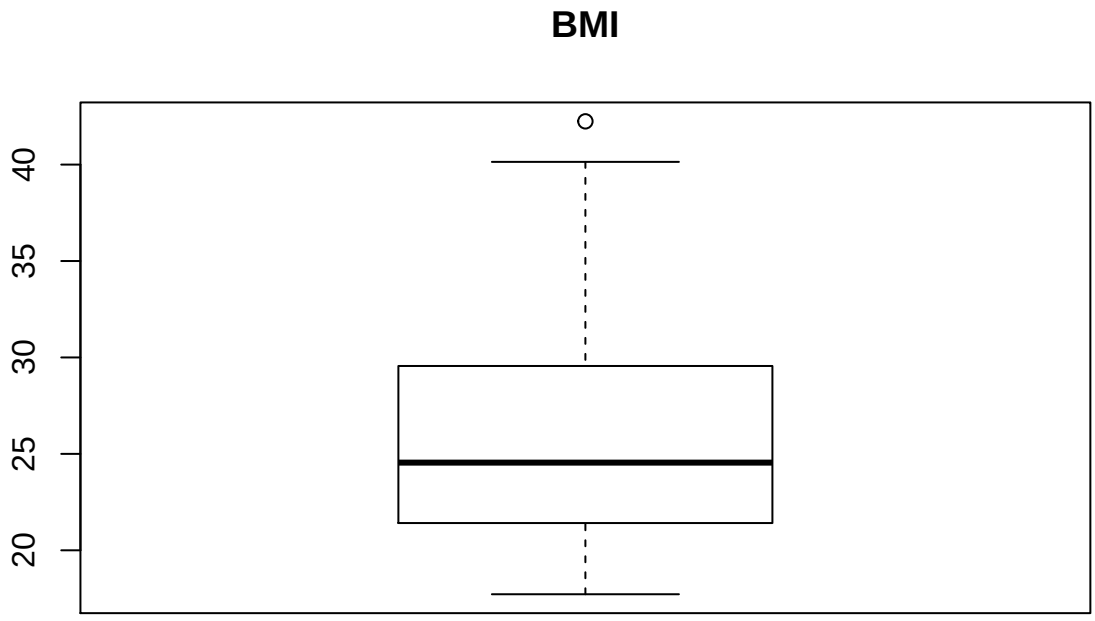
Se puede observar que existen valores que se han introducido de forma incorrecta, y explorando los datos, nos damos cuenta de que es una inconsistencia dada al introducir los datos, ya que el valor de *Anykids* es el mismo que el de *Numchild*. Resolveremos todas las inconsistencias vistas en este apartado en el punto 6.

```
data[which(data$Anykids > 1), ]
```

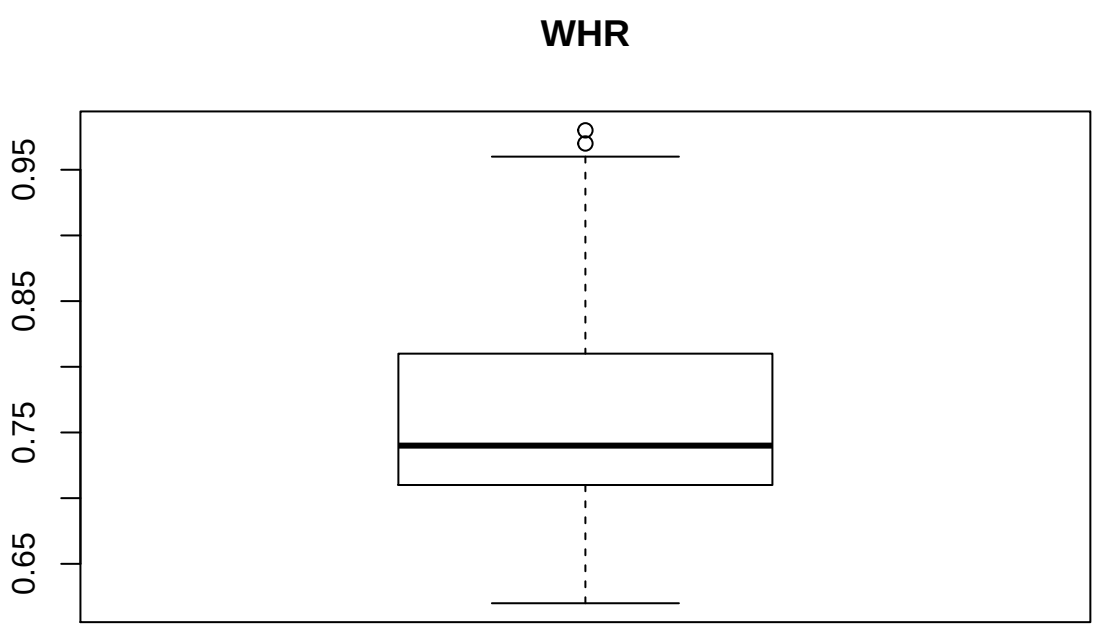
```
##       Id Estradl           Ethnic Entage Numchild Agefbo Anykids Agemenar
## 47  3300   12.60 African American     24        2     22       2     13.5
## 71    49   55.27        Caucasian     24        4     16       4     13.0
## 178 5016   43.30            Af Am     27        3     14       3     12.0
##         BMI  WHR Area
## 47  18.1384 0.74    0
## 71  18.8946 0.75    1
## 178 30.9264 0.86    0
```

En el caso de las variables *BMI* y *WHI*, que corresponden a la media de adiposidad general y la media de adiposidad abdominal respectivamente, obtenemos los siguientes diagramas de cajas, en los que se muestra que hay algún outlier. Sin embargo, puesto que dichos valores atípicos no distan mucho del máximo, los tomaremos como posibles valores.

```
boxplot(data$BMI,main="BMI")
```

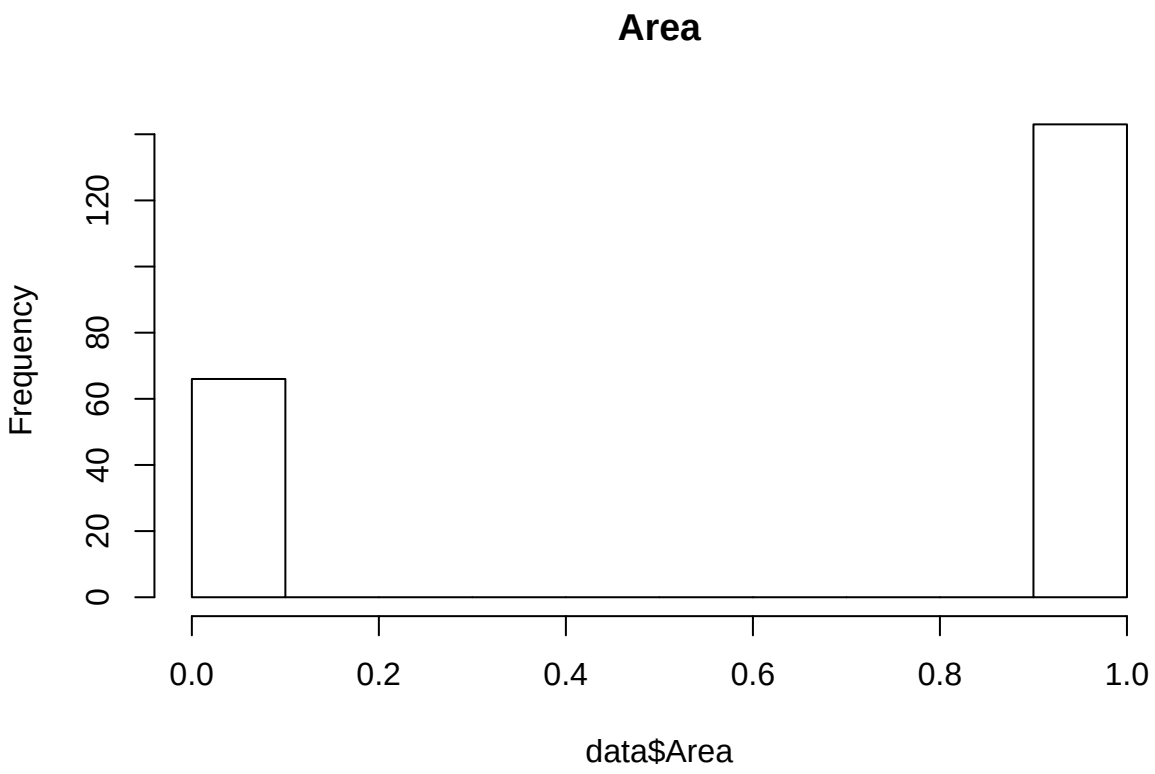



```
boxplot(data$WHR,main="WHR")
```



Por último, la variable area indica 0 si es rural o 1 si es urbana.

```
hist(data$Area, main="Area")
```



6. Inconsistencias

Las inconsistencias son errores que pueden haberse dado a causa de un fallo al introducir los datos, como puede ser que las variables *Anykids* y *Numchild* no estén relacionadas, indicando *Anykids* que la mujer no tiene hijos (valor = 0), mientras que *Numchild* indica que tiene uno o más, en cuyo caso prevalecerá el valor de *Numchild*. También puede ocurrir que la edad de la primera menarquía, *Agemenarq*, sea mayor que la edad actual, *Entage*, en cuyo caso intercambiaremos los valores. Debemos cerciorarnos también de que la edad en la que la mujer tuvo el primer hijo, *Agefbo*, sea superior a la edad de la primera menarquía, en caso contrario, indicaremos que la edad de la primera menarquía es 0.

En primer lugar, comprobamos si existe algun registro en el cual el valor de Anykids es *True* (no es 0), mientras que el valor *Numchild* indica que tiene menos de 1 hijo, y obtenemos que no exite dicha excepción, por lo que también comprobaremos la contraria con el mismo resultado.

```
which(data$Anykids > 1)
```

```
## [1]  46  70 176
```

```
which(data$Anykids > 0 & data$Numchild < 1)
```

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

```
which(data$Anykids == 0 & data$Numchild > 0)
```

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

Deberemos resolver también la inconsistencia que aparece cuando el valor de *Anykids* no es 0 o 1. Para ello, observamos los tres casos en los que ocurre la inconsistencia, dado que el valor *Anykids* ha sido introducido

de forma incorrecta, siendo este el mismo que el valor *Numchild*, y sustituimos su valor "booleano" por 1.

```
indexes <- which(data$Numchild > 0 & data$Anykids > 1)
data[indexes,]$Anykids <- 1
data[indexes, ]
```

```
##       Id Estradl           Ethnic Entage Numchild Agefbo Anykids Agemenar
## 47  3300   12.60 African American     24        2     22       1     13.5
## 71    49   55.27        Caucasian     24        4     16       1     13.0
## 178 5016   43.30            Af Am     27        3     14       1     12.0
##         BMI  WHR Area
## 47  18.1384 0.74    0
## 71  18.8946 0.75    1
## 178 30.9264 0.86    0
```

A continuación, observaremos la inconsistencia dada entre las variables *EntAge* y *Agemenarq*, que se da en un total de 9 registros.

```
index <- which(data$Agemenar > data$Entage)
data[index, ]
```

```
##       Id Estradl           Ethnic Entage Numchild Agefbo Anykids Agemenar
## 53  3460   12.90            Af Am     16        0      0       0       31
## 55  3500   29.80 African american     12       NA     NA      NA       22
## 126   48   49.50         Caucsian     11        0      0       0       25
## 152  228   18.37 African American     13        0      0       0       23
## 157 3440   15.30 African American     13        2     23       1       27
##         BMI  WHR Area
## 53  23.1901 0.73    1
## 55  19.8290 0.75    1
## 126 27.0953 0.76    1
## 152 27.8785 0.73    1
## 157 35.6201 0.75    1
```

```
real_agemenar <- data[which(data$Agemenar > data$Entage),]$Entage
real_entage <- data[which(data$Agemenar > data$Entage),]$Agemenar

data[index,]$Entage <- real_entage
data[index,]$Agemenar <- real_agemenar
data[index, ]
```

```
##       Id Estradl           Ethnic Entage Numchild Agefbo Anykids Agemenar
## 53  3460   12.90            Af Am     31        0      0       0       16
## 55  3500   29.80 African american     22       NA     NA      NA       12
## 126   48   49.50         Caucsian     25        0      0       0       11
## 152  228   18.37 African American     23        0      0       0       13
## 157 3440   15.30 African American     27        2     23       1       13
##         BMI  WHR Area
## 53  23.1901 0.73    1
## 55  19.8290 0.75    1
## 126 27.0953 0.76    1
## 152 27.8785 0.73    1
## 157 35.6201 0.75    1
```

Finalmente, la inconsistencia entre las variables *Agefbo* y *Agemenar* se pueden resolver como se muestra a continuación.

En primer lugar, observamos que existen registros que tienen la edad de la primera menarquía mayor que la

edad del primer hijo, por lo que anularemos uno de los dos valores asignándole el valor 0, dependiendo del valor *Numchild*. Es decir, si existe dicha inconsistencia y *Numchild* == 0, se anula la edad del primer hijo, en caso contrario, anularemos *Agemenar*.

```
which(data$Agemenar > data$Agefbo & data$Agefbo != 0)
```

```
## [1] 10
```

```
# Caso 1: Numchild == 0
which(data$Agemenar > data$Agefbo & data$Agefbo != 0 & data$Numchild == 0)
```

```
## [1] 10
```

Podemos comprobar que todos los casos en los que se da esta inconsistencia ocurren cuando el Número de hijos es 0, por lo que anularemos, asignando el valor 0 a *Agefbo* para dichos casos.

```
data[which(data$Agemenar > data$Agefbo & data$Agefbo != 0 & data$Numchild == 0), ]$Agefbo <- 0
data[which(data$Agefbo > 0 & data$Numchild == 0),]$Agefbo <- 0
```

7. Formato de las variables cualitativas

La variable cualitativa *Ethnic* únicamente puede tomar los valores "African American" y "Caucasian". Como podemos ver en la tabla de frecuencias que se muestra a continuación, esta variable no está estandarizada.

```
table(data$Ethnic)
```

```
##
##    African American                  Caucsian                  Af Am
##                  11                        11                    17
##       African american       African American              Caucacian
##                  22                       101                     3
##              Caucasian               Caucsian
##                  42                         2
```

Para estandarizar dicha variable, en primer lugar eliminaremos los espacios al inicio y final de cada cadena, para simplificar el ejercicio. Tenemos varias formas de proceder, buscar entre todos los datos una secuencia de caracteres común y asignar a todos los registros que encuentren dicha secuencia el valor correcto, como se ha realizado con el valor "African American", o tratar cada excepción de forma individual.

```
data$Ethnic <- trimws(data$Ethnic)

which(stringr::str_detect(data$Ethnic, "Af"))
```

```
##   [1]  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50
##  [18]  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67
##  [35]  68  71  72  73  74  75  82  83  84  85  86  88  89  90  97  98  99
##  [52] 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 117
##  [69] 118 120 121 122 123 125 126 127 128 129 130 131 133 134 135 136 137
##  [86] 138 139 140 141 142 143 144 145 147 148 149 150 151 152 153 154 155
## [103] 156 157 158 159 160 161 162 163 164 165 166 169 170 171 172 173 174
## [120] 175 176 177 178 179 180 181 182 183 184 185 189 190 191 192 193 194
## [137] 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209
```

```
data[which(stringr::str_detect(data$Ethnic, "Af")), ]$Ethnic <- "African American"

data[which(data$Ethnic == "Caucacian" | data$Ethnic == "Caucsian"), ]$Ethnic <- "Caucasian"
```

8. Imputación

Dados los valroes anómalos encontrados en el punto 5, procedemos a la imputación de los mismos aplicando la técnica de los k vecinos más cercanos (siendo k = 3), únicamente con los de su misma étnia, y utilizano la distancia de Gower. En primer lugar, guardamos los indices de todos los registros que contengan algún valor *NA*, que posteriormente vamos a imputar, para poder comprobar si los resultados son coherentes al final.

```
na_index <- which(is.na(data$Numchild) | is.na(data$Agefbo) | is.na(data$Anykids) | is.na(data$Agemenar
```

```
library(VIM)

index_af <- which(data$Ethnic == "African American")
data$Numchild[index_af] <- kNN(data[index_af,], variable = "Numchild", k = 3)$Numchild
data$Agefbo[index_af] <- kNN(data[index_af,], variable = "Agefbo", k = 3)$Agefbo
data$Anykids[index_af] <- kNN(data[index_af,], variable = "Anykids", k = 3)$Anykids
data$Agemenar[index_af] <- kNN(data[index_af,], variable = "Agemenar", k = 3)$Agemenar
```

```
index_ca <- which(data$Ethnic == "Caucasian")
data$Numchild[index_ca] <- kNN(data[index_ca,], variable = "Numchild", k = 3)$Numchild
data$Agefbo[index_ca] <- kNN(data[index_ca,], variable = "Agefbo", k = 3)$Agefbo
data$Anykids[index_ca] <- kNN(data[index_ca,], variable = "Anykids", k = 3)$Anykids
data$Agemenar[index_ca] <- kNN(data[index_ca,], variable = "Agemenar", k = 3)$Agemenar
```

Mostramos el resultado de los registros que previamente contenían valores *NA*.

```
data[na_index, c(3, 4, 5, 6, 7, 8)]
```

```
##                Ethnic Entage Numchild Agefbo Anykids Agemenar
## 7           Caucasian     32        2     27       1       13
## 45  African American     20        0      0       0       12
## 48  African American     32        0      0       0       14
## 55  African American     22        1     17       1       12
## 64  African American     32        0      0       0       14
## 122 African American     22        1     20       1       11
## 129 African American     35        1     19       1       12
## 135 African American     22        0      0       0       12
## 138 African American     32        1     24       1       12
## 150 African American     24        1     23       1       11
## 168 African American     27        1     22       1       12
```

9. Tabla resumen de las variables cualitativas

```
table(data$Ethnic)
```

```
##
## African American        Caucasian
##              151               58
```

10. Tabla resumen de las variables cuantitativas

En este punto observaremos tablas resumen de las variables cuantitativas, con medidas robustas, como puede ser la mediana o no robustas como la media, ya que es susceptible frente a *outliers*.

10.1 Estradl

```
summary(data$Estradl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.2    23.0    36.8    37.4    49.5    97.4
```

10.2 Entage

```
summary(data$Entage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   21.00   26.00   26.13   31.00   37.00
```

10.3 Numchild

```
summary(data$Numchild)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.5407  1.0000  7.0000
```

10.4 Agefbo

```
summary(data$Agefbo)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   6.526  17.000  32.000
```

10.5 Agemenar

```
summary(data$Agemenar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   11.00   12.00   12.37   13.00   16.00
```

10.6 BMI

```
summary(data$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.72   21.41   24.55   25.83   29.56   42.24
```

10.7 WHR

```
summary(data$WHR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6200  0.7100  0.7400  0.7594  0.8100  0.9800
```

11. Grabar el archivo

```
write.csv(data, file=paste(current_working_directory,"/ESTRADL_clean.csv", sep = ""))
```