Activitat 1: Preprocessament de dades

Proposta de solució

Semestre 2019.2

Índex

Les autoritats sanitàries volen fer un estudi del pes en néixer dels nens i nenes nascuts a Espanya en l'últim any. Per això, han realitzat un mostreig de 300 nens i nenes nascuts a diferents hospitals del país. El conjunt de dades **BWn.csv** conté aquestes dades.

L'estructura del conjunt de dades és la següent. Cada fila correspon a un naixement i per a cada un, s'especifiquen les següents variables:

- Id: identificador numèric.
- HP: nom de l'hospital.
- City: nom de la ciutat on s'ha produït el naixement.

- Time: hora del naixement (valor entre 0 i 24).
- Day: dia de la setmana (valor entre 1 i 7).
- BW: pes al naixement.
- BPD: diàmetre biparietal (en mm), determinat per ultrasons, mesurat abans de néixer.
- AD: diàmetre abdominal (en mm), determinat per ultrasons, mesurat abans de néixer.
- Sex: sexe del nadó.
- Ge: setmanes de gestació.
- Sm: si la mare és fumadora ('S' en cas afirmatiu, 'N' en cas negatiu).

L'objectiu d'aquesta activitat és preparar el fitxer pel seu posterior anàlisi. Per això, s'examinaran les dades de l'arxiu a la recerca de possibles errors, inconsistències i valors perduts i es realitzarà el preprocés adequat en cada cas.

Els criteris a aplicar per aquest preprocés són:

- Els noms de les ciutats comencen amb una lletra majúscula i van seguits per lletres minúscules.
- Els noms dels hospitals estandarditzats són els que apareixen a la llista que es mostra més endavant. Qualsevol variació sobre aquests noms ha de normalitzar-se.
- Es proporciona a més una llista dels hospitals i de les ciutats als quals corresponen. En cas de inconstistències, preval la dada de l'hospital i per tant, la ciutat es modifica per adequar-se.
- En variables quantitatives contínues, el símbol per expressar la coma decimal és el punt.
- El temps, expressat com un valor numèric de 0 a 24, s'ha d'expressar com HH:MM. No cal guardar el valor dels segons.
- Els dies de la setmana han de ser: Dilluns, Dimarts, Dimecres, Dijous, Divendres, Dissabte, Diumenge. Al fitxer, el número 1 correspon a Dilluns i consecutivament fins al 7 que correspon a Diumenge. Han de canviar-se els valors numèrics pel nom del dia.
- El pes (variable BW) ha d'expressar-se de forma numèrica i en grams. Per exemple, els valors del tipus "3.300gr" o "3.3Kg" han de guardar-se com una dada numèrica de valor 3300.
- Els valors de diàmetre biparietal i abdominal s'han d'expressar en mil · límetres i les variables han de ser numèriques. En cas que el format de les dades siguin de tipus caràcter, ha de transformar-se a numèric pel seu posterior anàlisi.
- El sexe ha de codificar com "F" per a referir-se a femení i "M" masculí.
- Les setmanes de gestació han d'arrodonir-se al valor enter més pròxim.
- En cas de realitzar canvis sobre les dades, aquests han de ser identificats i explicats en l'informe.

Els hospitals i les seves corresponents ciutats on s'han recollit les dades es mostren a la taula 1.

Un aspecte important del preprocessat és que s'han de documentar tots els canvis realitzats sobre el fitxer. En les anàlisis posteriors, s'han de conèixer els criteris de preprocessat i de quina manera han afectat el fitxer. És possible que alguns canvis introduïts (per exemple, una correcció sobre valors extrems) hagin de ser reconsiderats a la llum de les anàlisis posteriors. Per aquest motiu, no és suficient fer el preprocessat de les dades, sinó també documentar els canvis introduïts.

A continuació, s'especifiquen els passos a seguir per preprocessar el fitxer. Per millor sistematització de l'anàlisi, s'han de seguir els passos que s'indiquen en el mateix ordre i usant la mateixa numeració en els apartats.

Taula 1: Llistat d'hospitals i ciutats

Hospitales	Ciudades
H.U.La Paz	Madrid
H.Clínic	Barcelona
H.U.Vall d'Hebron	Barcelona
H.U.12 de octubre	Madrid
H.G.U.Gregorio Marañón	Madrid
H.U.Politècnic La Fe	Valencia
H.U.Quirón Dexeus	Barcelona
C.U.Navarra	Pamplona
H.U.de Bellvitge	Barcelona
H.M.I. Virgen de las Nieves	Granada
H.U.Virgen del Rocío	Sevilla

1 Càrrega de l'arxiu de dades i breu descripció

Obrir el fitxer **BWn.csv** i examinar el tipus de dades amb què R ha interpretat cada variable. Avaluar també els valors resum de cada tipus de variable.

```
#FUNCIÓ PER DOCUMENTAR ELS CANVIS INTRODUïTS EN EL PREPROCESSAT
report <- function( ds, row=0, message=""){
       i <- nrow(ds)-1
       rw <- data.frame(id=i+1, row, message)
       ds <- rbind( ds, rw )

     return (ds)
}
```

```
#LECTURA DEL FITXER
info <- data.frame(id=0, row=0, message="" )

ds <- read.csv(file="BWn.csv", sep=";")

class(ds)
```

```
## [1] "data.frame"
```

```
summary(ds)
```

```
##        ID                                    HP             City
##  Min.   :  1.00   H.U.de Bellvitge           : 40   Barcelona:124
##  1st Qu.: 75.75   H.U.Quirón Dexeus          : 36   Madrid   : 71
##  Median :150.50   H.U.Politècnic La Fe       : 31   Valencia : 31
##  Mean   :150.50   H.M.I. Virgen de las Nieves: 30   Granada  : 28
##  3rd Qu.:225.25   C.U.Navarra                : 27   Pamplona : 27
##  Max.   :300.00   H.G.U.Gregorio Marañón     : 27   Sevilla  : 14
##                   (Other)                    :109   (Other)  :  5
##       Time           Day               BW           BPD            AD
##  24     : 11   Min.   :1.000   2500  gr: 15   92 mm  : 37   105 mm : 29
##  6,84   :  3   1st Qu.:2.000   1700  gr: 11   91 mm  : 26   110 mm : 20
##  7,27   :  3   Median :4.000   2900  gr: 11   93 mm  : 25   90 mm  : 20
##  7,4    :  3   Mean   :4.043   2050  gr:  9   94 mm  : 25   108 mm : 19
```

```
##  9,35   :  3   3rd Qu.:6.000   2600  gr:  9   90 mm  : 24   100 mm : 17
##  10,4   :  2   Max.   :7.000   2750  gr:  9   (Other):153   (Other):190
##  (Other):275                   (Other) :236   NA's   : 10   NA's   :  5
##    Sex            Ge     Sm
##  boy : 14   33     :58   N:238
##  f   :  5   34     :48   S: 62
##  F   :149   37,5   :43
##  fem :  8   35,5   :36
##  girl:  6   43,2   :28
##  M   :118   32,8   :22
##             (Other):65
```

```
colnames(ds)
```

```
##  [1] "ID"   "HP"   "City" "Time" "Day"  "BW"   "BPD"  "AD"   "Sex"  "Ge"
## [11] "Sm"
```

```
id.factor   <- c(2,3,5,9,11)
id.num      <- c(6,7,8,10)
var.factor  <- colnames(ds)[id.factor]
var.num     <- colnames(ds)[id.num]
```

Haurien de ser variables qualitatives (factor): HP, City, Day, Sex, Sm

Haurien de ser variables quantitatives (numèriques): BW, BPD, AD, Ge

---

2 Normalització de les variables qualitatives

2.1 Hospital, Ciutat

En primer lloc, s'examinaran les possibles inconsistències en les variables HP i City. Detecteu si hi ha diversos noms per a un mateix hospital i/o per a una mateixa ciutat i si és així, normalitzar els noms, segons la llista de noms proporcionada anteriorment.

Comproveu així mateix la possible inconsistència entre l'hospital i la seva ciutat. En cas de incosistencia, identificar-la, reportar-la i introduir els canvis necessaris.

S'han de seguir els criteris especificats pel preprocessat.

```
  #Revisió dades hospitals
  table( ds$HP )
```

```
##
##                 C.U.Navarra                    H.Clínic
##                          27                          25
##     H.G.U.Gregorio Marañón H.M.I. Virgen de las Nieves
##                          27                          30
##          H.U.12 de octubre           H.U.de Bellvitge
##                          26                          40
##                 H.U.La Paz       H.U.Politècnic La Fe
##                          18                          31
##          H.U.Quirón Dexeus          H.U.Vall d'Hebron
```

```
##                          36                          26
##        H.U.Virgen del Rocío
##                          14
```

```
  #sense canvis

  #Revisió nom ciutats
  class( ds$City )
```

```
## [1] "factor"
```

```
  table( ds$City )
```

```
##
## Barcelona   Granada    Madrid  Pamplona   sevilla   Sevilla  valencia  Valencia
##       124        28        71        27         1        14         4        31
```

```
  ds$City <- as.character( ds$City )

  #Reporting de canvis
  idx <- which( ds$City=="sevilla")
  idx
```

```
## [1] 7
```

```
  info <- report(info, row=paste(idx,collapse=","), "Nom ciutat sevilla  -> Sevilla")
  idx <- which( ds$City=="valencia")
  idx
```

```
## [1]  1 20 27 29
```

```
  info <- report(info, row=paste(idx,collapse=","), "Nom ciutat valencia -> Valencia")
  #Canvis en el nom de la ciutat
  ds$City <- str_to_title(ds$City)
  ds$City <- as.factor( ds$City )
  levels(ds$City)
```

```
## [1] "Barcelona" "Granada"   "Madrid"    "Pamplona"  "Sevilla"   "Valencia"
```

```
  #Revisió d'inconsistencies Hosp-Ciutat
  #Taula correcta hospitals-noms
  Hosp.names
```

```
##                     Hospitales  Ciudades
## 1                   H.U.La Paz    Madrid
## 2                     H.Clínic Barcelona
## 3            H.U.Vall d'Hebron Barcelona
## 4            H.U.12 de octubre    Madrid
## 5       H.G.U.Gregorio Marañón    Madrid
## 6         H.U.Politècnic La Fe  Valencia
## 7            H.U.Quirón Dexeus Barcelona
## 8                  C.U.Navarra  Pamplona
## 9             H.U.de Bellvitge Barcelona
## 10 H.M.I. Virgen de las Nieves   Granada
## 11        H.U.Virgen del Rocío   Sevilla
```

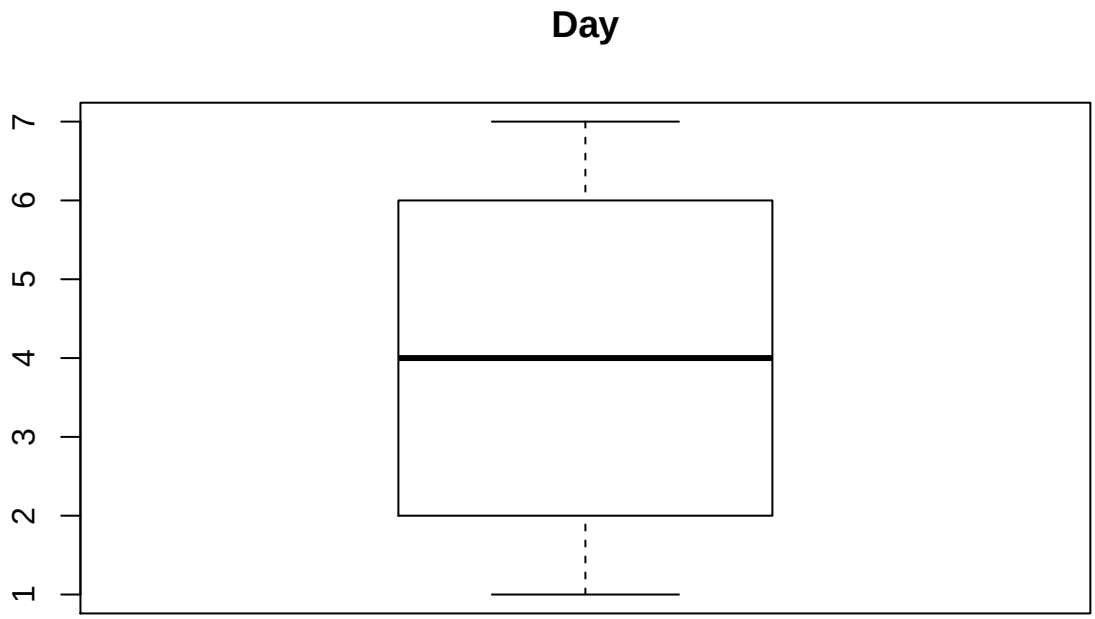
```
  i<-1
  for (i in 1:nrow(ds)){
     hosp <- ds$HP[i]
```

```
    city <- ds$City[i]
    idx <- which( as.character(Hosp.names$Hospitales)==as.character(hosp) )
    if ( as.character(Hosp.names$Ciudades[idx]) != as.character(city)){
      ds$City[i] <- Hosp.names$Ciudades[idx]
      info <- report( info, row=i, paste( "Inconsistencia hospital-ciutat: ",
                                          "Hosp=", hosp, ", ciutat:",  city, "->",
                                          Hosp.names$Ciudades[idx]) )
    }
  }
```

2.2 Dia de la setmana

Verificar que els valors de la variable dia són correctes i normalitzar la variable dia segons les indicacions proporcionades. La variable ha de ser de tipus categòric (en R, factor).

```
  boxplot( ds$Day, main="Day" )
```




```
  names.day <- c("Dilluns", "Dimarts", "Dimecres", "Dijous", "Divendres", "Dissabte", "Diumenge")
  class( ds$Day )
```

```
## [1] "integer"
```

```
  table( ds$Day )
```

```
##
##  1  2  3  4  5  6  7
```

```
## 33 43 45 41 59 53 26
```

```
  ds$Day <- as.factor( as.character( ds$Day ) )
  levels( ds$Day )
```

```
## [1] "1" "2" "3" "4" "5" "6" "7"
```

```
  info <- report( info, row="*", paste( "Dia de la setmana: {",
                                        paste(levels(ds$Day), collapse=', '),
                                        "} ->{" ,
                                        paste(names.day, collapse=', '),
                                        "}" ))
  levels( ds$Day ) <- names.day
  table( ds$Day )
```

```
##
##   Dilluns   Dimarts  Dimecres    Dijous Divendres  Dissabte  Diumenge
##        33        43        45        41        59        53        26
```

2.3 Sexe

Normalitzar la variable sexe (Sex) segons les indicacions proporcionades.

```
  table( ds$Sex )
```

```
##
##  boy    f    F  fem girl    M
##   14    5  149    8    6  118
```

```
  Sex <- ifelse( ds$Sex=="f" | ds$Sex=="fem" | ds$Sex=="girl" | ds$Sex=="F", "F", "M" )
  Sex<-as.factor(Sex)
  table(Sex)
```

```
## Sex
##   F   M
## 168 132
```

```
  info <- report( info, row="*", paste( "Variable Sexe: {",
                                        paste(as.character( levels(ds$Sex)), collapse=','),
                                        "} -> {",
                                         paste(as.character(levels(Sex)), collapse=','),
                                        "}"))
  ds$Sex <- as.factor(Sex)
```

2.4 Mare fumadora

Normalitzar els valors de la variable mare fumadora (Sm) segons les indicacions proporcionades.

```
table( ds$Sm )
```

```
##
##   N   S
## 238  62
```

```
#Sense canvis
```

3 Normalització de les variables quantitatives

Revisar el tipus de dada i el format de les variables que han de ser quantitatives. Convertiu a tipus numèric si les variables no s'han carregat amb aquest tipus. Abans, però, cal corregir les possibles inconsistències en el punt decimal. Reviseu, per a cada variable quantitativa el format especificat en els criteris de preprocés.

3.1 Pes

Tranformar la variable pes per convertir-la en un format numèric en grams. Per extreure el valor numèric d'un string en R, podeu fer servir expressions regulars i la funció str_extract. Podeu trobar informació a: https://stringr.tidyverse.org/articles/regular-expressions.html

```
   class(ds$BW)
```

```
## [1] "factor"
```

```
   head(ds$BW)
```

```
## [1] 2.2  kg  4100  gr 4.2  kg  1300  gr 1150  gr 3400  gr
## 91 Levels: 1.45  kg 1.65  kg 1.7  kg 1.804  kg 1025  gr 1150  gr ... 4850  gr
```

```
   #Canvi la coma decimal pel punt
   id <- grep( ",", ds$BW)
   ds$BW <- gsub( ",", ".", ds$BW)
   head(ds$BW)
```

```
## [1] "2.2  kg"  "4100  gr" "4.2  kg"  "1300  gr" "1150  gr" "3400  gr"
```

```
   info <- report( info, row=paste(id,collapse=','), message="BW: punt decimal")

   #Conversió a grams i a format numèric
   ikg <- grep( "kg", ds$BW )
   ig  <- grep( "gr", ds$BW)
   length(ikg) + length(ig)
```

```
## [1] 300
```

```
   ds$BW[ikg] <- str_extract( ds$BW[ikg], "\\-*\\d+\\.*\\d*")
   ds$BW[ig ] <- str_extract( ds$BW[ig], "\\-*\\d+\\.*\\d*")
   ds$BW <- as.numeric( ds$BW )
   ds$BW[ikg] <- ds$BW[ikg]*1000

   head(ds$BW)
```

```
## [1] 2200 4100 4200 1300 1150 3400
```

```
   class(ds$BW)
```

```
## [1] "numeric"
```

```
   info <- report( info, row="*", message=paste( "BW: kg->gr, files: ", paste(ikg,collapse=", ")) )
```

3.2 Diàmetre biparietal i Diàmetre abdominal

Normalitzar les variables diàmetre biparietal i diàmetre abdominal segons les indicacions proporcionades.

```
  class(ds$BPD)
```

```
## [1] "factor"
```

```
  summary(ds$BPD)
```

```
## 100 mm  64 mm  68 mm  72 mm  74 mm  80 mm  81 mm  82 mm  83 mm  84 mm  85 mm
##      3      3      6      1      5     13      2      7      8     12      6
##  86 mm  87 mm  88 mm  89 mm  90 mm  91 mm  92 mm  93 mm  94 mm  95 mm  96 mm
##      8     11     19     17     24     26     37     25     25     14      8
##  97 mm   NA's
##     10     10
```

```
  ds$BPD <- trimws( str_remove(ds$BPD, "mm") )
  id <- grep( ",", ds$BPD)
  id
```

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

```
  ds$BPD<-as.numeric( ds$BPD )

  info <- report( info, row="*", message="S'elimina \"mm\" de la variable BPD")

  class(ds$AD)
```

```
## [1] "factor"
```

```
  summary(ds$AD)
```

```
## 100 mm 101 mm 102 mm 103 mm 104 mm 105 mm 106 mm 107 mm 108 mm 109 mm 110 mm
##     17      6     15      9      4     29      6     10     19      2     20
## 112 mm 113 mm 114 mm 115 mm 116 mm 118 mm 119 mm 120 mm 129 mm 133 mm  71 mm
##      5      3      3      3      4      9      3      2      4      5      5
##  73 mm  78 mm  79 mm  80 mm  82 mm  85 mm  86 mm  89 mm  90 mm  92 mm  93 mm
##      1      6      1      5      5      3      4      6     20      8      8
##  94 mm  95 mm  96 mm  97 mm  98 mm  99 mm   NA's
##      8      7      8      5     14      3      5
```

```
  ds$AD <- trimws( str_remove(ds$AD, "mm") )
  id <- grep( ",", ds$AD)
  id
```

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

```
  ds$AD<-as.numeric( ds$AD )

  info <- report( info, row="*", message="S'elimina \"mm\" de la variable AD")
```

3.3 Setmanes de gestació

Normalitzar la variable setmanes de gestació segons els criteris establerts.

```
  class(ds$Ge)
```

```
## [1] "factor"
```

```
  summary(ds$Ge)
```

```
##  32,8    33 33,73    34  35,5 37,29  37,5  39,2  39,9  41,3  43,2
```

```
##    22    58    11    48    36    21    43    16     7    10    28
```

```
  idd<-grep( ",", ds$Ge)
  ds$Ge <- gsub( ",", ".", ds$Ge)
  info <- report( info, row="*", message=paste( "Ge: punt decimal. Files= ", paste(idd,collapse=", ")))

  #Redondeamos
  ds$Ge<- round( as.numeric( as.character(ds$Ge) ), 0 )
```

3.4 Hora

Transformar la variable hora a format HH: MM. Podeu fer servir llibreries o realitzar la transformació a partir del vostre propi codi.

```
str( ds$Time)
```

```
##  Factor w/ 267 levels "0","1,87","10,04",..: 81 25 162 143 265 179 252 190 221 63 ...
```

```
ds$Time <- gsub( ",", ".", as.character( ds$Time ))
T<-as.numeric( ds$Time )

#Examples: https://stackoverflow.com/questions/19721145/convert-hours-to-hhmmss-r
Tn<-format(ISOdatetime(1900,1,1,0,0,0, tz="GMT") + as.difftime(T, unit="hours"), "%H:%M")

#Validació
head(T)
```

```
## [1] 13.34 11.00 24.00 18.64  9.89  5.68
```

```
head(Tn)
```

```
## [1] "13:20" "11:00" "00:00" "18:38" "09:53" "05:40"
```

```
ds$Time <- Tn

info <- report( info, row="*", "Hora transformada en format HH:MM")
```

---

4 Valors perduts

Analitzar la presència de valors perduts. En el cas de detectar algun valor perdut en les variables quantitatives realitzar una imputació de valors en aquestes variables. La imputació s'ha de fer amb els 3 veïns més propers usant la distància de Gower, usant només la informació de les variables quantitatives i dins d'aquestes, aquelles que tinguin sentit en la imputació de la variable. Després de realitzar la imputació cal verificar que els valors assignats s'han copiat sobre el conjunt de dades originals. Visualitzar el resultat de les imputacions realitzades (per evitar mostrar tot el conjunt de dades, només s'han de mostrar els registres del conjunt de dades que contenen la imputació realitzada).

```
 idx <- !complete.cases(ds)
 sum(idx)
```

```
## [1] 15
```

Taula 2: Casos no complets

	ID	HP	City	Time	Day	BW	BPD	AD	Sex	Ge	Sm
24	24	H.U.Quirón Dexeus	Barcelona	10:33	Diumenge	3600	NA	116	F	41	N
53	53	H.U.La Paz	Madrid	18:28	Dimecres	2600	92	NA	F	34	N
65	65	H.U.La Paz	Madrid	09:31	Dimecres	2950	NA	107	M	38	N
101	101	H.U.de Bellvitge	Barcelona	01:52	Divendres	1250	NA	94	F	33	S
114	114	H.U.12 de octubre	Madrid	13:16	Divendres	3300	94	NA	F	39	N
127	127	H.U.de Bellvitge	Barcelona	21:24	Divendres	1150	68	NA	F	33	S
159	159	H.U.La Paz	Madrid	03:39	Dijous	2600	NA	98	F	34	N
183	183	H.U.12 de octubre	Madrid	09:32	Dimecres	4200	NA	133	F	43	N
186	186	H.U.12 de octubre	Madrid	11:04	Diumenge	1650	NA	82	F	33	S
192	192	H.G.U.Gregorio Marañón	Madrid	15:04	Dissabte	2700	NA	110	M	36	N
202	202	H.U.de Bellvitge	Barcelona	13:08	Dissabte	2200	82	NA	M	33	N
218	218	C.U.Navarra	Pamplona	11:55	Divendres	2500	NA	95	M	34	N
220	220	H.G.U.Gregorio Marañón	Madrid	12:13	Dilluns	2900	NA	110	F	38	N
225	225	H.Clínic	Barcelona	07:49	Dissabte	3000	NA	105	F	38	N
242	242	H.G.U.Gregorio Marañón	Madrid	09:21	Dimecres	1450	80	NA	F	33	S

```
info <- report(info, row="*",
               message= paste( sum(idx), "casos no complets en les files: ",
                               paste(which(idx==TRUE),collapse=", ") ) )

ds[idx,] %>% kable(caption="Casos no complets")
```

```
#Valors perduts en AD y BPD. Realitzem imputació amb Ge, BW, AD, BPD.
#La resta de variables s'asumeixen no relevants per la imputació
id.mis.bpd <- which( is.na(ds$BPD))
id.mis.ad  <- which( is.na(ds$AD))

#Imputació
selected.vars <- c("AD", "BPD", "BW","Ge")
output <- kNN( ds[,selected.vars], variable=c("BPD","AD"), k=3 )
output[output$BPD_imp==TRUE,]
```

```
##      AD BPD   BW Ge BPD_imp AD_imp
## 24  116  84 3600 41    TRUE  FALSE
## 65  107  92 2950 38    TRUE  FALSE
## 101  94  74 1250 33    TRUE  FALSE
## 159  98  85 2600 34    TRUE  FALSE
## 183 133  97 4200 43    TRUE  FALSE
## 186  82  85 1650 33    TRUE  FALSE
## 192 110  89 2700 36    TRUE  FALSE
## 218  95  90 2500 34    TRUE  FALSE
## 220 110  93 2900 38    TRUE  FALSE
## 225 105  91 3000 38    TRUE  FALSE
```

```
output[output$AD_imp==TRUE,]
```

```
##      AD BPD   BW Ge BPD_imp AD_imp
## 53  105  92 2600 34   FALSE   TRUE
## 114 110  94 3300 39   FALSE   TRUE
## 127  80  68 1150 33   FALSE   TRUE
```

```
## 202  90  82 2200 33   FALSE   TRUE
## 242  78  80 1450 33   FALSE   TRUE
```

```
 #Dataset després de la imputació
ds[,c("BPD","AD")] <- output[,c("BPD","AD")]
ds[ output$BPD_imp==TRUE | output$AD_imp==TRUE, ]
```

```
##      ID                    HP      City  Time       Day   BW BPD  AD Sex Ge Sm
## 24   24      H.U.Quirón Dexeus Barcelona 10:33  Diumenge 3600  84 116   F 41  N
## 53   53             H.U.La Paz    Madrid 18:28  Dimecres 2600  92 105   F 34  N
## 65   65             H.U.La Paz    Madrid 09:31  Dimecres 2950  92 107   M 38  N
## 101 101       H.U.de Bellvitge Barcelona 01:52 Divendres 1250  74  94   F 33  S
## 114 114      H.U.12 de octubre    Madrid 13:16 Divendres 3300  94 110   F 39  N
## 127 127       H.U.de Bellvitge Barcelona 21:24 Divendres 1150  68  80   F 33  S
## 159 159             H.U.La Paz    Madrid 03:39    Dijous 2600  85  98   F 34  N
## 183 183      H.U.12 de octubre    Madrid 09:32  Dimecres 4200  97 133   F 43  N
## 186 186      H.U.12 de octubre    Madrid 11:04  Diumenge 1650  85  82   F 33  S
## 192 192 H.G.U.Gregorio Marañón    Madrid 15:04  Dissabte 2700  89 110   M 36  N
## 202 202       H.U.de Bellvitge Barcelona 13:08  Dissabte 2200  82  90   M 33  N
## 218 218            C.U.Navarra  Pamplona 11:55 Divendres 2500  90  95   M 34  N
## 220 220 H.G.U.Gregorio Marañón    Madrid 12:13   Dilluns 2900  93 110   F 38  N
## 225 225               H.Clínic Barcelona 07:49  Dissabte 3000  91 105   F 38  N
## 242 242 H.G.U.Gregorio Marañón    Madrid 09:21  Dimecres 1450  80  78   F 33  S
```

```
info <- report(info, row="*", message=paste( "Valors perduts imputats en BPD: ", paste(id.mis.bpd, colla
info <- report(info, row="*",  message=paste( "Valors perduts imputats en AD: ", paste(id.mis.ad, collap
```

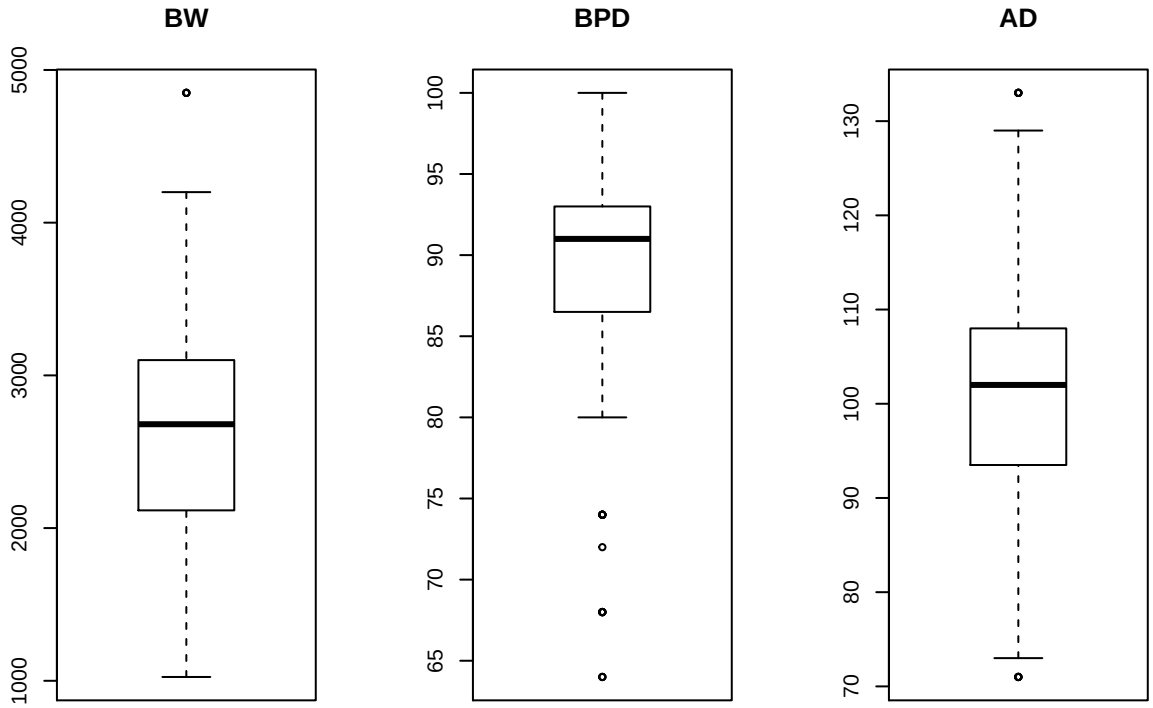
5 Valors extrems

Analitzar la presència de possibles valors extrems (outliers) en les variables pes, diàmetre biparietal i diàmetre abdominal. Per això, dibuixeu diagrames de caixa i també feu servir els resultats de la funció **boxplot.stats**. Un cop identificats, investigar les possibles causes d'aquests valors extrems i decidiu una estratègia apropiada, en funció d'aquesta anàlisi. L'estratègia pot ser eliminar els valors extrems, realitzar imputacions sobre els valors extrems o simplement mantenir els valors extrems pel seu valor explicatiu en el conjunt de dades. Justifiqueu les vostres eleccions.

```
par(mfrow=c(1,3))
boxplot( ds$BW, main="BW" )
boxplot( ds$BPD, main="BPD" )
boxplot( ds$AD, main="AD" )
```

Taula 3: Outliers en BW

	ID	HP	City	Time	Day	BW	BPD	AD	Sex	Ge	Sm
133	133	H.U.12 de octubre	Madrid	13:34	Dimecres	4850	95	118	M	43	N
151	151	H.Clínic	Barcelona	04:36	Dijous	4850	95	118	M	43	N
199	199	H.U.de Bellvitge	Barcelona	13:06	Diumenge	4850	95	118	F	43	N
264	264	H.U.Virgen del Rocío	Sevilla	11:33	Dilluns	4850	95	118	F	43	N




```
#Valors extrems i les seves posicions en BW:
values <- boxplot.stats(ds$BW)$out
idx <- which( ds$BW %in% values)
idx
```

```
## [1] 133 151 199 264
```

```
cat("Valors extrems en BW:", toString(values), "\n" )
```

```
## Valors extrems en BW: 4850, 4850, 4850, 4850
```

```
BW.outliers <- ds[idx,]
BW.outliers %>% kable( caption="Outliers en BW")
```

```
#Valors extrems i les seves posicions en BPD:
values <- boxplot.stats(ds$BPD)$out
idx <- which( ds$BPD %in% values)
cat("Valors extrems en BW:", toString(values), "\n" )
```

Taula 4: Outliers en BPD

	ID	HP	City	Time	Day	BW	BPD	AD	Sex	Ge	Sm
4	4	H.Clínic	Barcelona	18:38	Dijous	1300	74	71	M	33	S
5	5	H.U.12 de octubre	Madrid	09:53	Dimecres	1150	68	80	F	33	S
16	16	H.Clínic	Barcelona	00:00	Dimecres	1250	74	94	F	33	S
63	63	H.U.de Bellvitge	Barcelona	15:32	Dijous	1025	64	71	M	33	S
78	78	H.U.12 de octubre	Madrid	09:10	Divendres	1150	68	80	F	33	S
95	95	H.M.I. Virgen de las Nieves	Granada	00:00	Dissabte	1250	64	71	F	33	S
101	101	H.U.de Bellvitge	Barcelona	01:52	Divendres	1250	74	94	F	33	S
127	127	H.U.de Bellvitge	Barcelona	21:24	Divendres	1150	68	80	F	33	S
134	134	H.U.Politècnic La Fe	Valencia	06:43	Dilluns	1150	68	80	M	33	S
140	140	H.U.Politècnic La Fe	Valencia	13:32	Dijous	1250	74	94	F	33	S
155	155	H.U.Vall d'Hebron	Barcelona	16:18	Dimarts	1150	68	80	M	33	S
194	194	H.Clínic	Barcelona	12:40	Divendres	1250	64	71	M	33	S
210	210	C.U.Navarra	Pamplona	07:06	Dissabte	1300	74	71	M	33	S
240	240	H.U.de Bellvitge	Barcelona	08:22	Dissabte	1173	72	73	F	33	S
248	248	H.U.Quirón Dexeus	Barcelona	08:04	Dilluns	1150	68	80	F	33	S
286	286	H.U.Politècnic La Fe	Valencia	00:00	Divendres	1250	74	94	M	33	S

Taula 5: Outliers en AD

	ID	HP	City	Time	Day	BW	BPD	AD	Sex	Ge	Sm
3	3	H.U.de Bellvitge	Barcelona	00:00	Dimarts	4200	97	133	F	43	N
4	4	H.Clínic	Barcelona	18:38	Dijous	1300	74	71	M	33	S
47	47	H.U.Quirón Dexeus	Barcelona	14:16	Dissabte	4200	97	133	F	43	N
63	63	H.U.de Bellvitge	Barcelona	15:32	Dijous	1025	64	71	M	33	S
95	95	H.M.I. Virgen de las Nieves	Granada	00:00	Dissabte	1250	64	71	F	33	S
173	173	H.U.Vall d'Hebron	Barcelona	07:19	Dijous	4200	97	133	M	43	N
183	183	H.U.12 de octubre	Madrid	09:32	Dimecres	4200	97	133	F	43	N
194	194	H.Clínic	Barcelona	12:40	Divendres	1250	64	71	M	33	S
196	196	H.Clínic	Barcelona	14:00	Dijous	4200	97	133	F	43	N
210	210	C.U.Navarra	Pamplona	07:06	Dissabte	1300	74	71	M	33	S

```
## Valors extrems en BW: 74, 68, 74, 64, 68, 64, 74, 68, 68, 74, 68, 64, 74, 72, 68, 74
```

```
BPD.outliers <- ds[idx,]
BPD.outliers %>% kable( caption="Outliers en BPD")
```

```
#Valors extrems i les seves posicions en AD:
values <- boxplot.stats(ds$AD)$out
idx <- which( ds$AD %in% values)
cat("Valors extrems en BW:", toString(values), "\n" )
```

```
## Valors extrems en BW: 133, 71, 133, 71, 71, 133, 133, 71, 133, 71
```

```
AD.outliers <- ds[idx,]
AD.outliers %>% kable( caption="Outliers en AD")
```

Interpretació:

- Els valors extrems en BW corresponen a casos de pes baix (per sota de 1200gr) i alguns casos de pes elevat (4850gr).

Taula 6: Estadística descriptiva de variables qualitatives

	HP	City	Day	Sex	Sm
	H.U.de Bellvitge : 40	Barcelona:127	Dilluns :33	F:168	N:238
	H.U.Quirón Dexeus : 36	Granada : 30	Dimarts :43	M:132	S: 62
	H.U.Politècnic La Fe : 31	Madrid : 71	Dimecres :45		
	H.M.I. Virgen de las Nieves: 30	Pamplona : 27	Dijous :41		
	C.U.Navarra : 27	Sevilla : 14	Divendres:59		
	H.G.U.Gregorio Marañón : 27	Valencia : 31	Dissabte :53		
	(Other) :109		Diumenge :26		

- Els casos extrems en BPD (diàmetre biparietal) corresponen a valors baixos (entre 68mm i 74mm) i tots ells corresponen a valors de baix pes al néixer.
- Els casos extrems en AD corresponen a baix diàmetre abdominal (menys de 73mm) o a valors elevats (més de 120 mm). En el primer cas, són valors de pes baix, mentre que el segon tipus correspon a valors elevats de pes.

Per tant, s'observa una relació entre baix pes i baix diàmetre abdominal i biparietal, i a l'inversa. Sense més informació sobre l'anàlisi, es prefereix no modificar aquests valors del conjunt de dades ja que poden aportar informació útil a l'anàlisi (per exemple, si els casos de baix pes corresponen amb dones fumadores o no).

6 Taula resum de les variables qualitatives

Realitzar un resum descriptiu dels valors de les variables qualitatives.

```
#var.factor  <- c("HP", "City","Day","Sex", "Sm")
options(knitr.kable.NA = '')
kable( summary(ds)[,id.factor],
      digits=2, align='l', caption="Estadística descriptiva de variables qualitatives")
```

7 Taula resum de les variables quantitatives

Realitzar una taula de la tendència central i dispersió de les variables quantitatives. Feu servir mesures robustes i no robustes.

```
#var.num     <- c("ID", "Time", "BW", "BPD", "AD", "Ge")
mean.n <- as.vector(sapply( ds[,id.num ],mean,na.rm=TRUE ) )
std.n <- as.vector(sapply( ds[,id.num ],sd, na.rm=TRUE))
median.n <- as.vector(sapply( ds[,id.num],median, na.rm=TRUE))
mean.trim.0.05 <- as.vector(sapply( ds[,id.num],mean, na.rm=TRUE, trim=0.05))
mean.winsor.0.05 <- as.vector(sapply( ds[,id.num],winsor.mean, na.rm=TRUE,trim=0.05))
IQR.n <- as.vector(sapply( ds[,id.num],IQR, na.rm=TRUE))
mad.n <- as.vector(sapply( ds[,id.num],mad, na.rm=TRUE))

kable(data.frame(variables= names( ds)[id.num],
Media = mean.n,
Mediana = median.n,
Media.recort.0.05= mean.trim.0.05,
Media.winsor.0.05= mean.winsor.0.05
```

Taula 7: Estimacions de Tendència Central

variables	Media	Mediana	Media.recort.0.05	Media.winsor.0.05
BW	2660.79	2680	2650.15	2650.13
BPD	88.94	91	89.55	89.09
AD	100.97	102	100.97	100.77
Ge	36.24	36	36.04	36.24

Taula 8: Estimacions de Dispersió

variables	SD	IQR	MAD
BW	760.91	976.00	696.82
BPD	6.35	6.25	4.45
AD	11.66	14.25	10.38
Ge	3.17	5.00	2.97

```
  ),
  digits=2, caption="Estimacions de Tendència Central")
```

```
  kable(data.frame(variables=names(ds)[id.num],
    SD = std.n,
    IQR = IQR.n,
    MAD = mad.n),
  digits=2, caption="Estimacions de Dispersió")
```

8 Crear el fitxer net

Graveu les dades preprocessades en un fitxer anomenat "BWprocessed.csv".

```
my.newfile <- "BWprocessed.csv"
write.csv(ds, file=my.newfile, row.names = FALSE)
```

9 Documentació

Documentar de forma resumida els canvis introduïts en el fitxer durant el preprocés d'ell mateix. Cal mostrar-se en forma de taula, indicant el tipus de preprocés aplicat en cada cas. Cal explicar el detall del preprocés aplicat. Per exemple, no és suficient dir "s'ha normalitzat la variable BW". En tot cas, s'hauria d'indicar si s'ha reemplaçat la coma pel punt decimal, o si s'han arrodonit decimals, etcètera. Heu de ser específics, ja que l'informe ha de ser útil com a documentació dels canvis realitzats.

10 Comentaris importants

1. **No es pot inspeccionar ni corregir de manera manual** el fitxer de dades. Per exemple, **no** es pot fer una assignació del tipus:

`` ` data[1,5] <- 32.5` ``

Taula 9: Resum de preprocessament (quan s'especifica * vol dir que s'aplica a diversos registres o tots del conjunt de dades)

id	row	message
1	7	Nom ciutat sevilla -> Sevilla
2	1,20,27,29	Nom ciutat valencia -> Valencia
3	1	Inconsistencia hospital-ciutat: Hosp= H.G.U.Gregorio Marañón , ciutat: Valencia -> Madrid
4	3	Inconsistencia hospital-ciutat: Hosp= H.U.de Bellvitge , ciutat: Madrid -> Barcelona
5	7	Inconsistencia hospital-ciutat: Hosp= H.M.I. Virgen de las Nieves , ciutat: Sevilla -> Granada
6	15	Inconsistencia hospital-ciutat: Hosp= H.M.I. Virgen de las Nieves , ciutat: Madrid -> Granada
7	20	Inconsistencia hospital-ciutat: Hosp= H.Clínic , ciutat: Valencia -> Barcelona
8	27	Inconsistencia hospital-ciutat: Hosp= H.U.La Paz , ciutat: Valencia -> Madrid
9	29	Inconsistencia hospital-ciutat: Hosp= H.U.de Bellvitge , ciutat: Valencia -> Barcelona
10	*	Dia de la setmana: { 1, 2, 3, 4, 5, 6, 7 } ->{ Dilluns, Dimarts, Dimecres, Dijous, Divendres, Dissabte, Diumenge }
11	*	Variable Sexe: { boy,f,F,fem,girl,M } -> { F,M }
12	36,43,45	BW: punt decimal
13	*	BW: kg->gr, files: 1, 3, 14, 17, 23, 24, 27, 36, 43, 45, 55, 59, 60, 73, 76, 86, 89, 98, 105, 116, 117, 118, 125, 137, 149, 151, 152, 156, 157, 162, 165, 176, 183, 187, 198, 201, 207, 215, 218, 227, 231, 236, 246, 249, 254, 255, 257, 258, 268, 269, 277, 279, 283, 284, 287, 298
14	*	S'elimina "mm" de la variable BPD
15	*	S'elimina "mm" de la variable AD
16	*	Ge: punt decimal. Files= 1, 2, 3, 6, 8, 11, 12, 14, 15, 17, 18, 19, 23, 24, 27, 30, 32, 34, 35, 36, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 55, 57, 58, 59, 61, 64, 65, 66, 67, 69, 71, 73, 74, 76, 77, 80, 81, 82, 83, 84, 86, 89, 90, 91, 92, 94, 96, 98, 99, 100, 102, 104, 105, 106, 107, 108, 110, 112, 113, 114, 115, 116, 117, 118, 119, 120, 122, 124, 129, 130, 131, 132, 133, 135, 137, 138, 139, 144, 145, 147, 148, 149, 150, 151, 153, 154, 156, 158, 160, 161, 163, 164, 165, 166, 167, 168, 169, 170, 171, 173, 174, 176, 177, 178, 179, 181, 182, 183, 184, 185, 187, 188, 191, 192, 193, 195, 196, 197, 198, 199, 202, 203, 204, 206, 207, 208, 209, 211, 214, 215, 219, 220, 221, 222, 225, 227, 229, 230, 232, 233, 238, 239, 243, 244, 245, 246, 247, 249, 250, 251, 252, 253, 257, 259, 260, 262, 263, 264, 268, 269, 270, 271, 273, 275, 279, 280, 281, 282, 283, 287, 288, 289, 290, 291, 292, 293, 295, 298, 299, 300
17	*	Hora transformada en format HH:MM
18	*	15 casos no complets en les files: 24, 53, 65, 101, 114, 127, 159, 183, 186, 192, 202, 218, 220, 225, 242
19	*	Valors perduts imputats en BPD: 24,65,101,159,183,186,192,218,220,225
20	*	Valors perduts imputats en AD: 53,114,127,202,242

Aquest tipus de transformacions s'han de fer amb funcionalitats de búsqueda (cercar els registres que tenen errors o inconsistències) i després fer les correccions oportunes amb funcionalitats d'R. Així el procediment de neteja és útil, independentment del fitxer de dades i de la posició i valors concrets del fitxer.

2. **No es poden fer llistats complets de les dades del fitxer a pantalla**, perquè genera fitxers de sortida excessivament grans. Si voleu testejar el resultat d'una instrucció sobre les dades, podeu usar la funció **head** que mostra les primeres files de la taula de dades o **tail** que mostra les darreres.

11 Puntuacions dels apartats

- Seccions 1,2 (20%)
- Secció 3 (20%)
- Secció 4 (20%)
- Secció 5 (10%)
- Seccions 6,7 (10%)
- Seccions 8,9 (10%)
- Puntuacions dels apartats (10%)

12 Referències

Quick-R

Cookbook for R

LaTex tables