

TFM - Anexo de Prácticas

Alumno Román Octavio Calafati - Máster Bioinformática y Bioestadística - UOC

Fecha límite de entrega para la Memoria Final: Miércoles 24 de mayo de 2017 a las 12 p.m.



3.2.1 Visión práctica del problema

```
# Calculo la media
y <- c(4, 6, 7)
mean(y)

## [1] 5.666667

# Calculo la media, pero hay un dato faltante
y <- c(4, 6, NA)
mean(y)

## [1] NA

# Utilizando el argumento na.rm=TRUE
y <- c(4, 6, NA)
mean(y, na.rm=TRUE)

## [1] 5

# Verifico cual es el directorio de trabajo
getwd()

## [1] "/Users/MacMini/UOC - M0.168 TFM"

# Defino directorio de trabajo
setwd("~/UOC - M0.168 TFM")

# Quito todas las variables de ambiente
rm(list=ls())

# Primer vistazo de los contenidos del fichero
readLines("H0C.csv", n=5)
```

```
## [1]
"\",\", \"surgery\", \"Age\", \"Hospital_Number\", \"rectal_temp\", \"pulse\", \
\"resp_rate\", \"extremities_temp\", \"peripheral_pulse\", \"mucous_membranes
\", \"capillary_refill_time\", \"pain\", \"peristalsis\", \"abdominal_distens
ion\", \"nasogastric_tube\", \"nasogastric_reflux\", \"nasogastric_reflux_PH
\", \"rectal_examination_feces\", \"abdomen\", \"packed_cell_volume\", \"tota
l_protein\", \"abdominocentesis_appearance\", \"abdomcentesis_total_protein
\", \"outcome\", \"surgical_lesion\", \"type_of_lesion_1\", \"type_of_lesion_
2\", \"type_of_lesion_3\", \"cp_data\""
## [2]
"\1\", 2, 1, 530101, 38.5, 66, 28, 3, 3, NA, 2, 5, 4, 4, NA, NA, NA, 3, 5, 45, 8.4, NA, NA, 2, 2
, 11300, 0, 0, 2"
## [3]
"\2\", 1, 1, 534817, 39.2, 88, 20, NA, NA, 4, 1, 3, 4, 2, NA, NA, NA, 4, 2, 50, 85, 2, 2, 3, 2, 2
208, 0, 0, 2"
## [4]
"\3\", 2, 1, 530334, 38.3, 40, 24, 1, 1, 3, 1, 3, 3, 1, NA, NA, NA, 1, 1, 33, 6.7, NA, NA, 1, 2,
0, 0, 0, 1"
## [5]
"\4\", 1, 9, 5290409, 39.1, 164, 84, 4, 1, 6, 2, 2, 4, 4, 1, 2, 5, 3, NA, 48, 7.2, 3, 5.3, 2, 1,
2208, 0, 0, 1"
```

Ahora Leo el fichero .csv y lo cargo

```
HOC <- read.table("HOC.csv", sep=",", dec=".", header=TRUE,
stringsAsFactors=FALSE)
names(HOC)
```

```
## [1] "X" "surgery"
## [3] "Age" "Hospital_Number"
## [5] "rectal_temp" "pulse"
## [7] "resp_rate" "extremities_temp"
## [9] "peripheral_pulse" "mucous_membranes"
## [11] "capillary_refill_time" "pain"
## [13] "peristalsis" "abdominal_distension"
## [15] "nasogastric_tube" "nasogastric_reflux"
## [17] "nasogastric_reflux_PH" "rectal_examination_feces"
## [19] "abdomen" "packed_cell_volume"
## [21] "total_protein" "abdominocentesis_appearance"
## [23] "abdomcentesis_total_protein" "outcome"
## [25] "surgical_lesion" "type_of_lesion_1"
## [27] "type_of_lesion_2" "type_of_lesion_3"
## [29] "cp_data"
```

Verifico estructura del dataframe

```
str(HOC)
```

```
## 'data.frame': 300 obs. of 29 variables:
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...
## $ surgery : int 2 1 2 1 2 2 1 1 2 2 ...
## $ Age : int 1 1 1 9 1 1 1 1 1 9 ...
## $ Hospital_Number : int 530101 534817 530334 5290409
```

```

530255 528355 526802 529607 530051 5299629 ...
## $ rectal_temp : num 38.5 39.2 38.3 39.1 37.3 NA 37.9
NA NA 38.3 ...
## $ pulse : int 66 88 40 164 104 NA 48 60 80 90
...
## $ resp_rate : int 28 20 24 84 35 NA 16 NA 36 NA ...
## $ extremities_temp : int 3 NA 1 4 NA 2 1 3 3 1 ...
## $ peripheral_pulse : int 3 NA 1 1 NA 1 1 NA 4 NA ...
## $ mucous_membranes : int NA 4 3 6 6 3 1 NA 3 1 ...
## $ capillary_refill_time : int 2 1 1 2 2 1 1 1 1 1 ...
## $ pain : int 5 3 3 2 NA 2 3 NA 4 5 ...
## $ peristalsis : int 4 4 3 4 NA 3 3 4 4 3 ...
## $ abdominal_distension : int 4 2 1 4 NA 2 3 2 4 1 ...
## $ nasogastric_tube : int NA NA NA 1 NA 2 1 2 2 2 ...
## $ nasogastric_reflux : int NA NA NA 2 NA 1 1 1 1 1 ...
## $ nasogastric_reflux_PH : num NA NA NA 5 NA NA NA NA NA NA ...
## $ rectal_examination_feces : int 3 4 1 3 NA 3 3 3 3 3 ...
## $ abdomen : int 5 2 1 NA NA 3 5 4 5 NA ...
## $ packed_cell_volume : num 45 50 33 48 74 NA 37 44 38 40 ...
## $ total_protein : num 8.4 85 6.7 7.2 7.4 NA 7 8.3 6.2
6.2 ...
## $ abdominocentesis_appearance: int NA 2 NA 3 NA NA NA NA NA 1 ...
## $ abdomcentesis_total_protein: num NA 2 NA 5.3 NA NA NA NA NA 2.2
...
## $ outcome : int 2 3 1 2 2 1 1 2 3 1 ...
## $ surgical_lesion : int 2 2 2 1 2 2 1 1 1 2 ...
## $ type_of_lesion_1 : int 11300 2208 0 2208 4300 0 3124
2208 3205 0 ...
## $ type_of_lesion_2 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ type_of_lesion_3 : int 0 0 0 0 0 0 0 0 0 0 ...
## $ cp_data : int 2 2 1 1 2 2 2 2 2 1 ...

# Guardo el dataset HOC para generar el fichero HOC.rds
# (a efectos de reproducibilidad posterior)
saveRDS(HOC, "HOC.rds")

# Apertura del dataset HOC
# (Descomentar la línea siguiente si fuera necesario abrirlo para
reproducibilidad posterior)
# HOC <- readRDS("HOC.rds")

```

3.2.2 Listwise Deletion

```

# Genero modelo de regresión lineal
fit<-lm(pulse ~ type_of_lesion_1, data=HOC)

# Se borran los registros con datos faltantes al utilizar el argumento

```

```
'na.action = na.omit'
fit <- lm(pulse ~ type_of_lesion_1, data=HOC, na.action = na.omit)
coef(fit)

##      (Intercept) type_of_lesion_1
##      7.037461e+01      4.271436e-04

# Verificar cuantos registros fueron eliminados
eliminados <- na.action(fit)
naprint(eliminados)

## [1] "24 observations deleted due to missingness"

# Más registros con valores faltantes si agrego la variable
"abdominal_distension" como predictora
fit2 <- lm(pulse ~ type_of_lesion_1 + abdominal_distension, data=HOC)
naprint(na.action(fit2))

## [1] "71 observations deleted due to missingness"
```

3.2.3 Pairwise Deletion

```
colMeans(HOC, na.rm = TRUE)

##              X              surgery
##      1.505000e+02      1.397993e+00
##              Age      Hospital_Number
##      1.640000e+00      1.085889e+06
##      rectal_temp              pulse
##      3.816792e+01      7.191304e+01
##      resp_rate      extremities_temp
##      3.041736e+01      2.348361e+00
##      peripheral_pulse      mucous_membranes
##      2.017316e+00      2.853755e+00
##      capillary_refill_time      pain
##      1.305970e+00      2.951020e+00
##      peristalsis      abdominal_distension
##      2.917969e+00      2.266393e+00
##      nasogastric_tube      nasogastric_reflux
##      1.755102e+00      1.582474e+00
##      nasogastric_reflux_PH      rectal_examination_feces
##      4.707547e+00      2.757576e+00
##      abdomen      packed_cell_volume
##      3.692308e+00      4.629520e+01
##      total_protein      abdominocentesis_appearance
##      2.445693e+01      2.037037e+00
##      abdomcentesis_total_protein      outcome
##      3.019608e+00      1.551839e+00
##      surgical_lesion      type_of_lesion_1
```

```
##          1.363333e+00          3.657880e+03
##          type_of_lesion_2          type_of_lesion_3
##          9.022667e+01          7.363333e+00
##          cp_data
##          1.670000e+00

a <- cor(HOC, use = "pair")

# Visualizo sólo las 6 primeras filas
head(a)

##          X          surgery          Age Hospital_Number
## X          1.000000000 -0.09275015 -0.05348758 -0.009961148
## surgery    -0.092750153  1.000000000 -0.08931879 -0.126043799
## Age        -0.053487583 -0.08931879  1.000000000  0.697433955
## Hospital_Number -0.009961148 -0.12604380  0.69743396  1.000000000
## rectal_temp  0.014783271  0.04068010  0.19684404  0.144420722
## pulse       -0.049863144 -0.18688550  0.52983820  0.380078482
##          rectal_temp          pulse          resp_rate extremities_temp
## X          0.01478327 -0.04986314 -0.01735362  0.03070778
## surgery    0.04068010 -0.18688550 -0.20142323 -0.11647364
## Age        0.19684404  0.52983820  0.41227590 -0.05673518
## Hospital_Number 0.14442072  0.38007848  0.27397575 -0.06116699
## rectal_temp  1.00000000  0.21926017  0.26918132  0.09652555
## pulse       0.21926017  1.00000000  0.47039707  0.33504101
##          peripheral_pulse mucous_membranes
capillary_refill_time
## X          0.034791682 -0.070219369 -
0.10023742
## surgery    -0.238054131 -0.187695693 -
0.08742574
## Age        0.030656059 -0.006628693 -
0.01586790
## Hospital_Number 0.009016445 -0.031190330
0.04499351
## rectal_temp  0.093773772  0.051793906
0.13932397
## pulse       0.533369102  0.483060063
0.39848128
##          pain peristalsis abdominal_distension
## X          0.004712395 -0.11891225 -0.03506056
## surgery    -0.316086444 -0.28588936 -0.24713695
## Age        0.034857959 -0.03954913  0.09637696
## Hospital_Number 0.097597480 -0.01735768  0.06792270
## rectal_temp -0.088079389  0.04698099  0.04965014
## pulse       0.307522021  0.33530047  0.42333034
##          nasogastric_tube nasogastric_reflux
nasogastric_reflux_PH
## X          0.02455320  0.01452943
0.066097812
```

| | | | |
|--------------------|-----------------------------|-----------------------------|------------------|
| ## surgery | -0.15051000 | -0.12867267 | |
| 0.175800531 | | | |
| ## Age | 0.09218690 | -0.07147698 | |
| 0.028004041 | | | |
| ## Hospital_Number | -0.01610032 | -0.03093299 | - |
| 0.005577291 | | | |
| ## rectal_temp | -0.14263271 | 0.01446057 | |
| 0.228305910 | | | |
| ## pulse | 0.05980760 | 0.23402363 | |
| 0.005795796 | | | |
| ## | rectal_examination_feces | abdomen | |
| packed_cell_volume | | | |
| ## X | -0.06404337 | -0.001536589 | - |
| 0.09246395 | | | |
| ## surgery | -0.24886861 | -0.393516084 | - |
| 0.04203374 | | | |
| ## Age | -0.08634740 | -0.094955534 | - |
| 0.14741539 | | | |
| ## Hospital_Number | -0.09923886 | 0.008340622 | - |
| 0.09305766 | | | |
| ## rectal_temp | 0.02899817 | -0.037589873 | |
| 0.06537316 | | | |
| ## pulse | 0.26670718 | 0.274881266 | |
| 0.40609021 | | | |
| ## | total_protein | abdominocentesis_appearance | |
| ## X | 0.02619737 | 0.02574511 | |
| ## surgery | -0.04482801 | -0.21259132 | |
| ## Age | -0.11937987 | -0.07770150 | |
| ## Hospital_Number | -0.22978938 | -0.16010906 | |
| ## rectal_temp | -0.06225120 | -0.07826784 | |
| ## pulse | -0.09159872 | 0.36204924 | |
| ## | abdomcentesis_total_protein | outcome | |
| surgical_lesion | | | |
| ## X | -0.06034290 | 0.029697349 | - |
| 0.02524883 | | | |
| ## surgery | 0.03513254 | -0.117619377 | |
| 0.60502302 | | | |
| ## Age | -0.05876734 | -0.004083472 | - |
| 0.04393995 | | | |
| ## Hospital_Number | -0.04098076 | -0.004376518 | - |
| 0.10139308 | | | |
| ## rectal_temp | 0.01127046 | -0.057082020 | |
| 0.01955315 | | | |
| ## pulse | 0.03144616 | 0.313215253 | - |
| 0.26633548 | | | |
| ## | type_of_lesion_1 | type_of_lesion_2 | type_of_lesion_3 |
| ## X | 0.14238933 | -0.010672090 | 0.05308885 |
| ## surgery | -0.16394008 | -0.098594360 | -0.04710089 |
| ## Age | 0.02296642 | 0.017915483 | 0.19611614 |
| ## Hospital_Number | 0.14548573 | -0.050252610 | -0.02087546 |

```
## rectal_temp      0.06270311    -0.062728167    0.02939493
## pulse            0.07951079    -0.001454538    0.10146572
##                  cp_data
## X                0.007244350
## surgery          0.005826918
## Age              -0.080481522
## Hospital_Number -0.097457068
## rectal_temp      -0.066317467
## pulse            -0.119113065
```

```
b <- cov(HOC, use = "pair")
```

```
# Visualizo sólo las 6 primeras filas
head(b)
```

```
##                  X          surgery          Age
Hospital_Number
## X                7.525000e+03 -3.951224e+00 -1.008696e+01 -
1.321897e+06
## surgery          -3.951224e+00  2.403986e-01 -9.535140e-02 -
9.467934e+04
## Age              -1.008696e+01 -9.535140e-02  4.726154e+00
2.319487e+06
## Hospital_Number -1.321897e+06 -9.467934e+04  2.319487e+06
2.340291e+12
## rectal_temp      9.420607e-01  1.473401e-02  3.193863e-01
1.667279e+05
## pulse            -1.255605e+02 -2.631387e+00  3.292838e+01
1.590550e+07
##                  rectal_temp          pulse          resp_rate
extremities_temp
## X                9.420607e-01 -1.255605e+02 -2.675143e+01
2.767658e+00
## surgery          1.473401e-02 -2.631387e+00 -1.748651e+00 -
6.035098e-02
## Age              3.193863e-01  3.292838e+01  1.605487e+01 -
1.176550e-01
## Hospital_Number 1.667279e+05  1.590550e+07  7.482344e+06 -
9.241454e+04
## rectal_temp      5.362467e-01  4.621599e+00  3.553247e+00
7.495770e-02
## pulse            4.621599e+00  8.197088e+02  2.440669e+02
9.899851e+00
##                  peripheral_pulse mucous_membranes
capillary_refill_time
## X                3.125711e+00    -9.828330e+00    -
4.130080e+00
## surgery          -1.226505e-01    -1.503194e-01    -2.058517e-
02
## Age              6.113307e-02     -2.095489e-02     -1.520487e-
```

```

02
## Hospital_Number      1.288242e+04      -6.927985e+04
3.083278e+04
## rectal_temp          6.764690e-02      6.435709e-02      4.614570e-
02
## pulse                1.524186e+01      2.213819e+01
5.194071e+00
##                      pain    peristalsis abdominal_distension
## X                    5.284209e-01 -1.007999e+01      -3.199622e+00
## surgery              -2.041928e-01 -1.368689e-01      -1.298847e-01
## Age                  9.287387e-02 -7.916667e-02      1.976658e-01
## Hospital_Number      1.775179e+05 -2.265648e+04      1.045943e+05
## rectal_temp          -7.944724e-02  3.422192e-02      3.637186e-02
## pulse                1.120930e+01  9.339338e+00      1.260839e+01
##                      nasogastric_tube nasogastric_reflux
nasogastric_reflux_PH
## X                    1.375092e+00      1.007211e+00
1.145831e+01
## surgery              -4.789849e-02      -5.124676e-02
1.666183e-01
## Age                  1.105181e-01      -8.610651e-02
1.184325e-01
## Hospital_Number      -1.440225e+04      -3.268155e+04      -
1.684099e+04
## rectal_temp          -6.786950e-02      8.977667e-03
3.552128e-01
## pulse                1.031962e+00      4.827542e+00
3.400000e-01
##                      rectal_examination_feces      abdomen
packed_cell_volume
## X                    -7.045993e+00 -2.035699e-01      -
8.392053e+01
## surgery              -1.530353e-01 -2.889503e-01      -
2.156409e-01
## Age                  -2.067374e-01 -1.665958e-01      -
3.430805e+00
## Hospital_Number      -1.659386e+05  1.217909e+04      -
1.450103e+06
## rectal_temp          2.451369e-02 -4.364802e-02
4.965554e-01
## pulse                8.991557e+00  1.021440e+01
1.189696e+02
##                      total_protein abdominocentesis_appearance
## X                    6.265164e+01      1.790492e+00
## surgery              -6.035438e-01      -7.889126e-02
## Age                  -7.228645e+00      -1.481481e-01
## Hospital_Number      -9.504629e+06      -1.989880e+05
## rectal_temp          -1.279962e+00      -5.349985e-02
## pulse                -7.172957e+01      8.362392e+00
##                      abdomcentesis_total_protein      outcome

```



```

surgical_lesion
## X -1.009390e+01 1.902157e+00 -
1.055184e+00
## surgery 3.232673e-02 -4.251307e-02
1.430159e-01
## Age -2.500485e-01 -6.554286e-03 -
4.602007e-02
## Hospital_Number -1.037371e+05 -4.942812e+03 -
7.472689e+04
## rectal_temp 1.949566e-02 -3.068458e-02
7.029289e-03
## pulse 1.795176e+00 6.726277e+00 -
3.694704e+00
## type_of_lesion_1 type_of_lesion_2 type_of_lesion_3
## X 6.669378e+04 -6.013512e+02 5.873428e+02
## surgery -4.347362e+02 -3.145263e+01 -2.950226e+00
## Age 2.695888e+02 2.529926e+01 5.437538e+01
## Hospital_Number 1.201738e+09 -4.993660e+07 -4.072922e+06
## rectal_temp 2.347719e+02 -3.207523e+01 3.069339e+00
## pulse 1.213215e+04 -2.818277e+01 3.862694e+02
## cp_data
## X 2.959866e-01
## surgery 1.346771e-03
## Age -8.240803e-02
## Hospital_Number -7.022111e+04
## rectal_temp -2.339435e-02
## pulse -1.614862e+00

```

3.2.4 Imputación de la media

```

# Cargo Librería MICE
library("mice")

# Visualizo el patrón de datos faltantes
c <- md.pattern(HOC)

# Visualizo sólo las 6 primeras filas
head(c)

## X Age Hospital_Number surgical_lesion type_of_lesion_1
type_of_lesion_2
## 6 1 1 1 1 1
1
## 1 1 1 1 1 1
1
## 1 1 1 1 1 1
1

```

```

## 15 1 1 1 1 1 1
1
## 1 1 1 1 1 1
1
## 5 1 1 1 1 1
1
## type_of_lesion_3 cp_data surgery outcome pulse packed_cell_volume
## 6 1 1 1 1 1 1
## 1 1 1 1 1 1
## 1 1 1 1 1 1
## 15 1 1 1 1 1
## 1 1 1 1 1 1
## 5 1 1 1 1 1
## capillary_refill_time total_protein peristalsis mucous_membranes
pain
## 6 1 1 1 1
1
## 1 1 1 1
1
## 1 1 1 1
1
## 15 1 1 1
1
## 1 1 1 1
1
## 5 1 1 1
1
## extremities_temp abdominal_distension resp_rate rectal_temp
## 6 1 1 1 1
## 1 1 1 0 1
## 1 1 1 1 1
## 15 1 1 1 1
## 1 1 1 1 1
## 5 1 1 1 1
## peripheral_pulse rectal_examination_feces nasogastric_tube
## 6 1 1 1
## 1 1 1 1
## 1 1 1 0
## 15 1 1 1
## 1 1 0 1
## 5 1 1 1
## nasogastric_reflux abdomen abdominocentesis_appearance
## 6 1 1 1
## 1 1 1 1
## 1 1 1 1
## 15 1 1 1
## 1 1 1 1
## 5 1 0 1
## abdomcentesis_total_protein nasogastric_reflux_PH
## 6 1 1 0

```

```
## 1      1      1 1
## 1      1      1 1
## 15     1      0 1
## 1      1      1 1
## 5      1      1 1

# Generación de subsample mediante paquete DPLYR
library("dplyr")
subsample <- slice(HOC,1:10) %>% select(pulse, abdominal_distension, pain
, nasogastric_reflux)

# Visualización de subsample
subsample

##   pulse abdominal_distension pain nasogastric_reflux
## 1    66                   4    5                NA
## 2    88                   2    3                NA
## 3    40                   1    3                NA
## 4   164                   4    2                 2
## 5   104                  NA   NA                NA
## 6    NA                   2    2                 1
## 7    48                   3    3                 1
## 8    60                   2   NA                 1
## 9    80                   4    4                 1
## 10   90                   1    5                 1

# Genero el patrón de datos faltantes de subsample
pattern <- md.pattern(subsample)

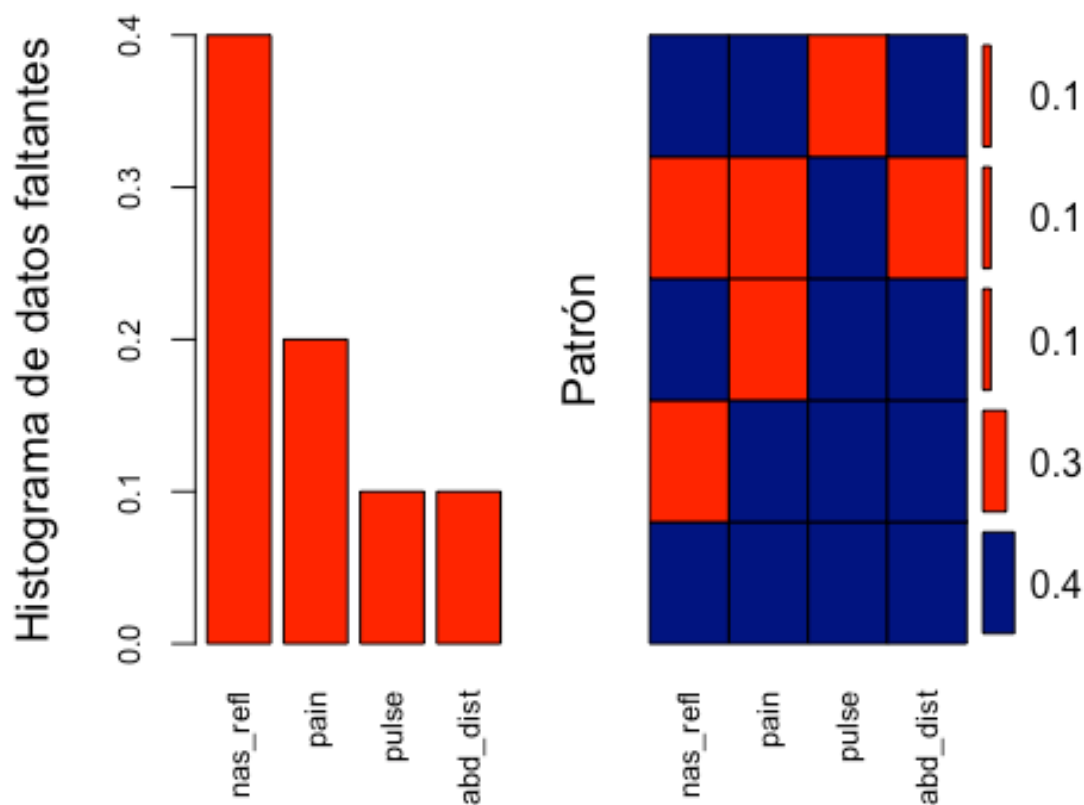
# Abrevio los nombres de las variables de subsample
names(subsample) <- c("pulse","abd_dist","pain", "nas_refl")

# Visualizo el patrón de datos faltantes de subsample
pattern

##   pulse abdominal_distension pain nasogastric_reflux
## 4     1                   1    1                1 0
## 1     0                   1    1                1 1
## 1     1                   1    0                1 1
## 3     1                   1    1                0 1
## 1     1                   0    0                0 3
##      1                   1    2                4 8

# Cargo librería VIM
library(VIM)

# Ilustración 12: Visualización de patrón de valores faltantes en
subsample, mediante paquete VIM
aggr_plot <- aggr(subsample, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(subsample), cex.axis=.8, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable Count
## nas_refl  0.4
##      pain  0.2
##      pulse 0.1
##      abd_dist 0.1

# Copia de seguridad del dataset HOC
HOC_bak <- HOC

# Imputar la media al dataset HOC
HOC_mean_imp <- mice(HOC, method="mean", m=1, maxit=1)

##
## iter imp variable
## 1 1 surgery rectal_temp pulse resp_rate extremities_temp
peripheral_pulse mucous_membranes capillary_refill_time pain
peristalsis abdominal_distension nasogastric_tube nasogastric_reflux
nasogastric_reflux_PH rectal_examination_feces abdomen
packed_cell_volume total_protein abdominocentesis_appearance
abdomcentesis_total_protein outcome
```

```

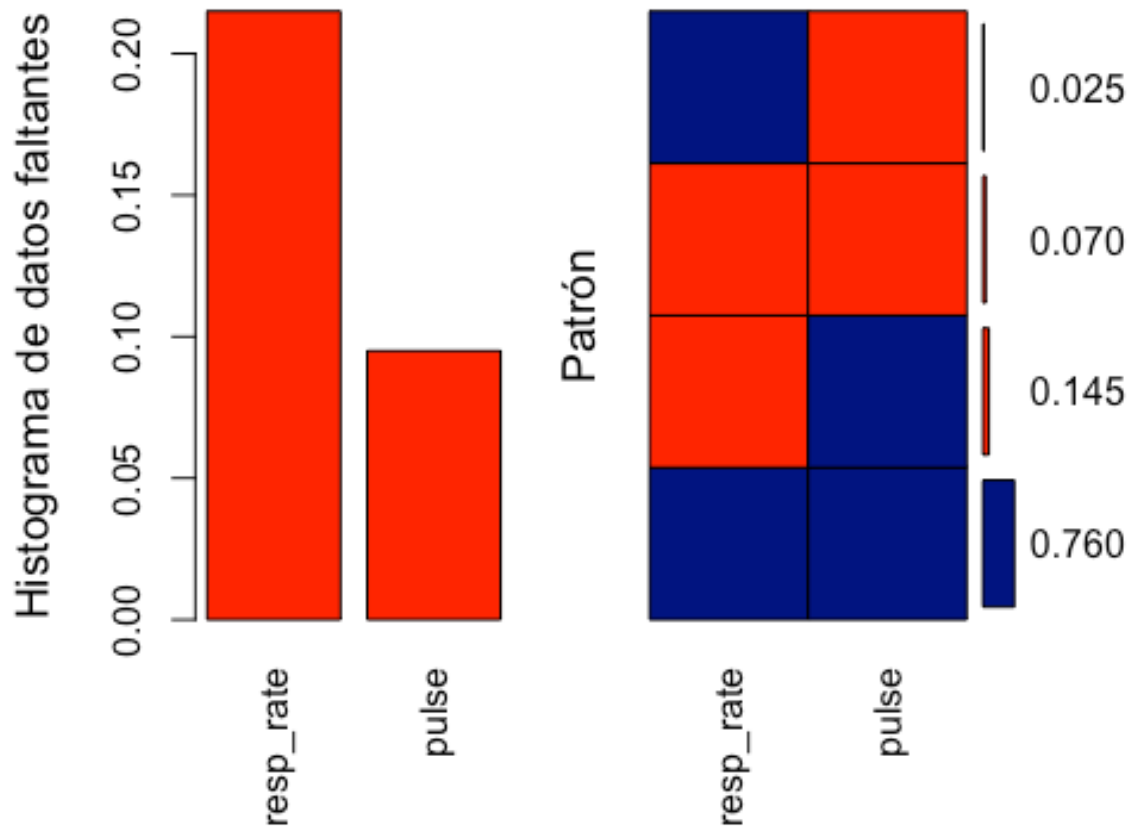
# Generación de subsample2 mediante paquete DPLYR
subsample2 <- slice(HOC,1:200) %>% select(pulse, resp_rate)

# Visualizo subsample2 (Sólo las 6 primeras filas)
head(subsample2)

##   pulse resp_rate
## 1    66        28
## 2    88        20
## 3    40        24
## 4   164        84
## 5   104        35
## 6    NA         NA

# Ilustración 13: Visualización de patrón de valores faltantes en
subsample2, mediante paquete VIM
aggr_plot <- aggr(subsample2, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(subsample2), cex.axis=1, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))

```



```

##
## Variables sorted by number of missings:
## Variable Count

```

```

## resp_rate 0.215
## pulse 0.095

# Imputar la media al dataset subsample2
imp <- mice(subsample2, method="mean", m=1, maxit=1)

##
## iter imp variable
## 1 1 pulse resp_rate

# Cargo Librería Lattice
library(lattice)

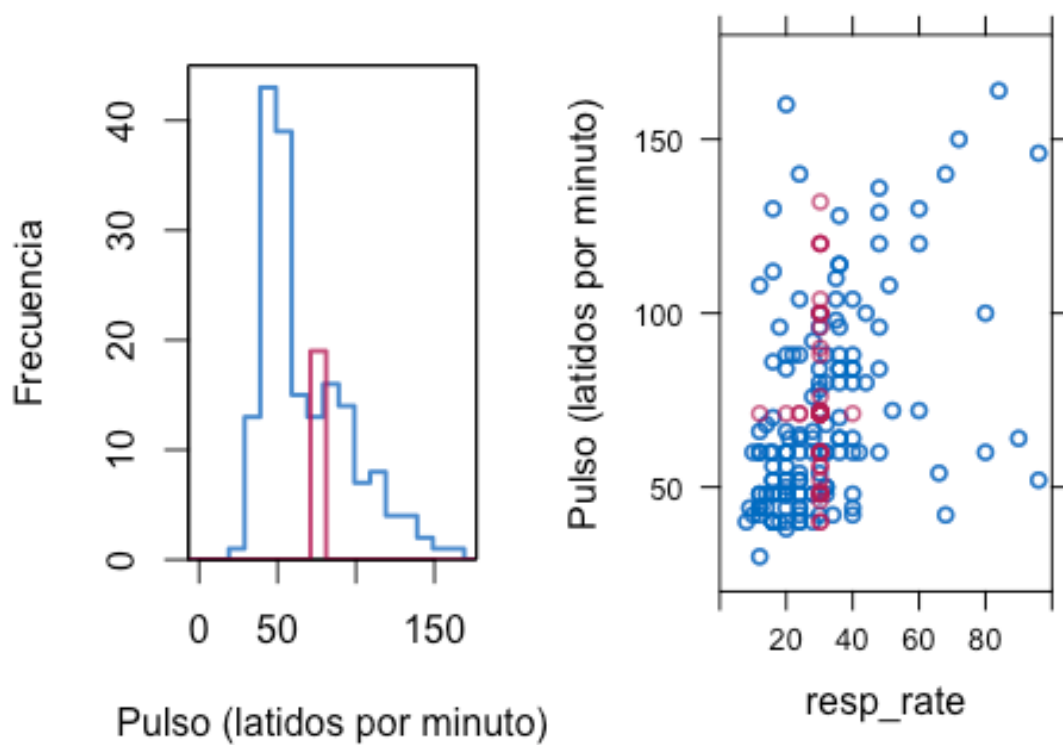
# Ilustración 14
lwd <- 1.5
par(mfrow=c(1,2))
breaks <- seq(-20, 200, 10)
nudge <- 1
x <- matrix(c(breaks-nudge, breaks+nudge), ncol=2)
obs <- subsample2[, "pulse"]
mis <- imp$imp$pulse[, 1]
fobs <- c(hist(obs, breaks, plot=FALSE)$counts, 0)
fmis <- c(hist(mis, breaks, plot=FALSE)$counts, 0)
y <- matrix(c(fobs, fmis), ncol=2)
matplot(x, y, type="s",
        col=c(mdc(4),mdc(5)), lwd=2, lty=1,
        xlim = c(0, 170), ylim = c(0,45), yaxs = "i",
        xlab="Pulso (latidos por minuto)",
        ylab="Frecuencia")
box()

tp <- xyplot(imp, pulse~resp_rate, na.groups=ici(imp),
            ylab="Pulso (latidos por minuto)", xlab="resp_rate",
            cex = 0.75, lex=lwd,
            ylim = c(20, 180), xlim = c(0,100))
print(tp, newpage = FALSE, position = c(0.48,0.045,1,0.95))

# Verificación del modelo de imputación empleado en cada variable
imp$meth

## pulse resp_rate
## "mean" "mean"

```



3.2.5 Imputación mediante regresión

```
# Defino modelo y predicciones para subsample2
fit <- lm(pulse ~ resp_rate, data=subsample2)
pred <- predict(fit, newdata=ic(subsample2))

# Imputación alternativa utilizando MICE
imp <- mice(subsample2[,1:2], method="norm.predict", m=1, maxit=3, seed=1)

##
## iter imp variable
## 1 1 pulse resp_rate
## 2 1 pulse resp_rate
## 3 1 pulse resp_rate

# Ilustración 15
par(mfrow = c(1,2))
fmis <- c(hist(pred, breaks, plot=FALSE)$counts, 0)
y <- matrix(c(fobs, fmis), ncol=2)
matplot(x, y, type="s",
```

```

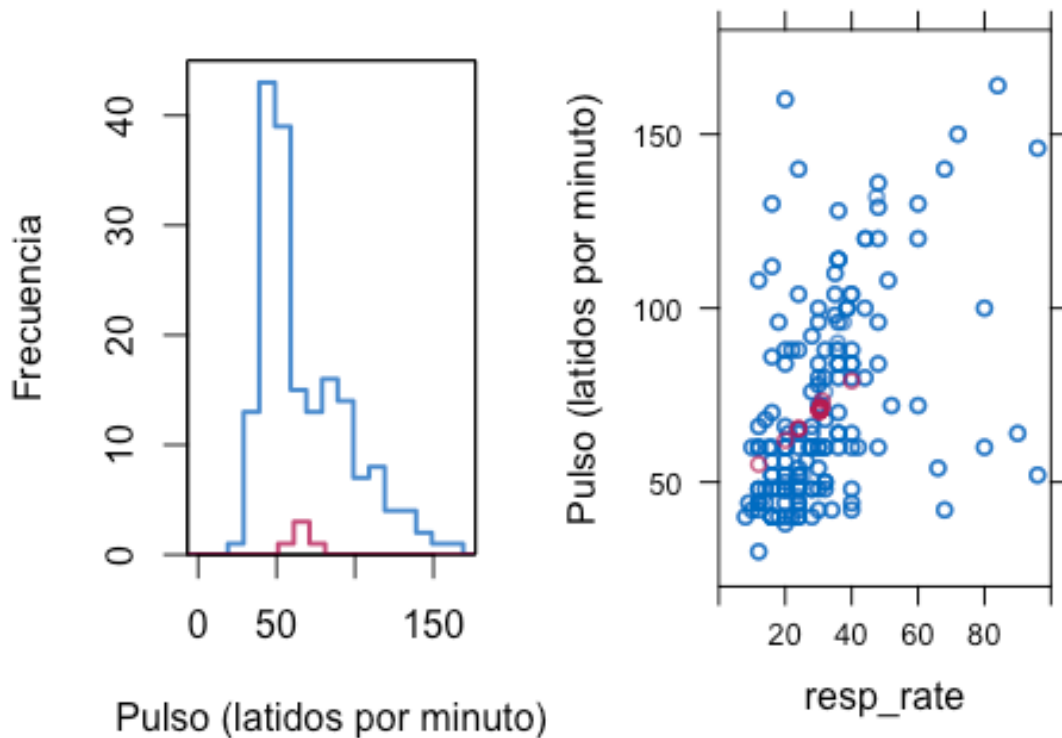
col=c(mdc(4),mdc(5)), lwd=2, lty=1,
xlim = c(0, 170), ylim = c(0,45), yaxs = "i",
xlab="Pulso (latidos por minuto)",
ylab="Frecuencia")
box()

tp <- xyplot(imp, pulse~resp_rate,
  ylab="Pulso (latidos por minuto)", xlab="resp_rate",
  cex = 0.75, lex=lwd,
  ylim = c(20, 180), xlim = c(0,100))
print(tp, newpage = FALSE, position = c(0.48,0.045,1,0.95))

# Verificación del modelo de imputación empleado en cada variable
imp$meth

##           pulse           resp_rate
## "norm.predict" "norm.predict"

```



3.2.6 Imputación mediante regresión estocástica

Realizo la imputación mediante regresión estocástica, utilizando el paquete MICE:

```
imp <- mice(subsample2[,1:2],method="norm.nob",m=1,maxit=1,seed=1)
```

```
##
```

```
## iter imp variable
```

```
## 1 1 pulse resp_rate
```

Ilustración 16

```
par(mfrow = c(1,2))
```

```
mis <- imp$imp$pulse[,1]
```

```
fmis <- c(hist(mis, breaks, plot=FALSE)$counts, 0)
```

```
y <- matrix(c(fobs, fmis), ncol=2)
```

```
matplot(x, y, type="s",
```

```
        col=c(mdc(4),mdc(5)), lwd=2, lty=1,
```

```
        xlim = c(0, 170), ylim = c(0,45), yaxs = "i",
```

```
        xlab="Pulso (latidos por minuto)",
```

```
        ylab="Frecuencia")
```

```
box()
```

```
tp <- xyplot(imp, pulse~resp_rate, na.groups=ici(imp),
```

```
            ylab="Pulso (latidos por minuto)", xlab="resp_rate",
```

```
            cex = 0.75, lex=lwd,
```

```
            ylim = c(20, 180), xlim = c(0,100))
```

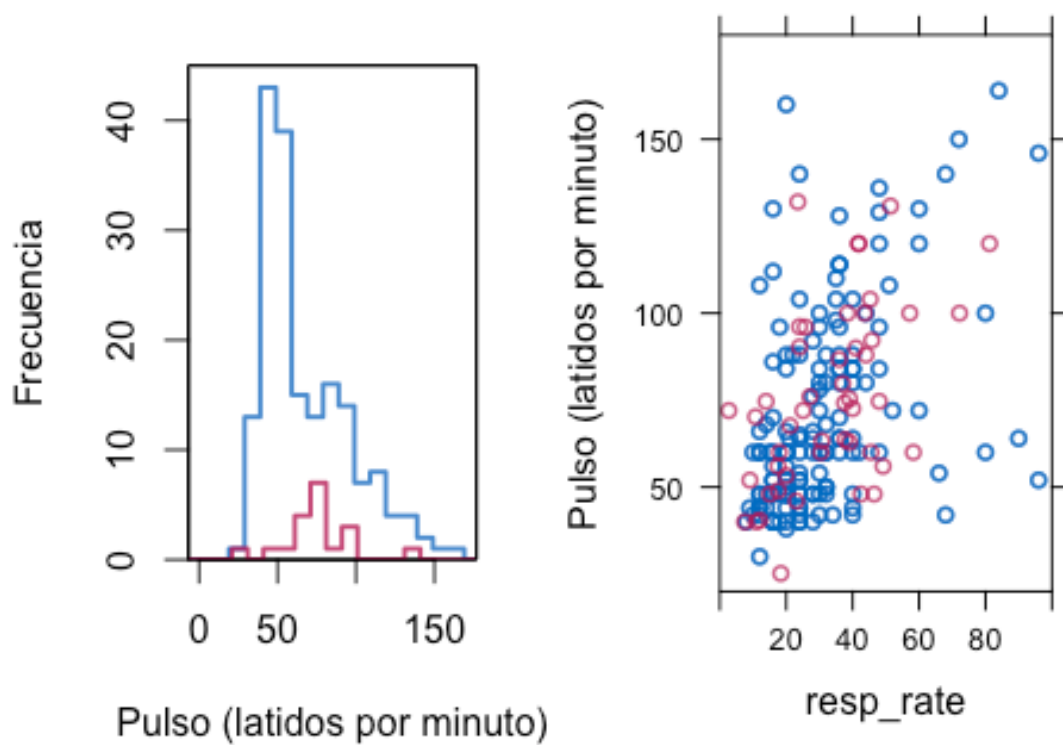
```
print(tp, newpage = FALSE, position = c(0.48,0.045,1,0.95))
```

Verificación del modelo de imputación empleado en cada variable

```
imp$meth
```

```
## pulse resp_rate
```

```
## "norm.nob" "norm.nob"
```

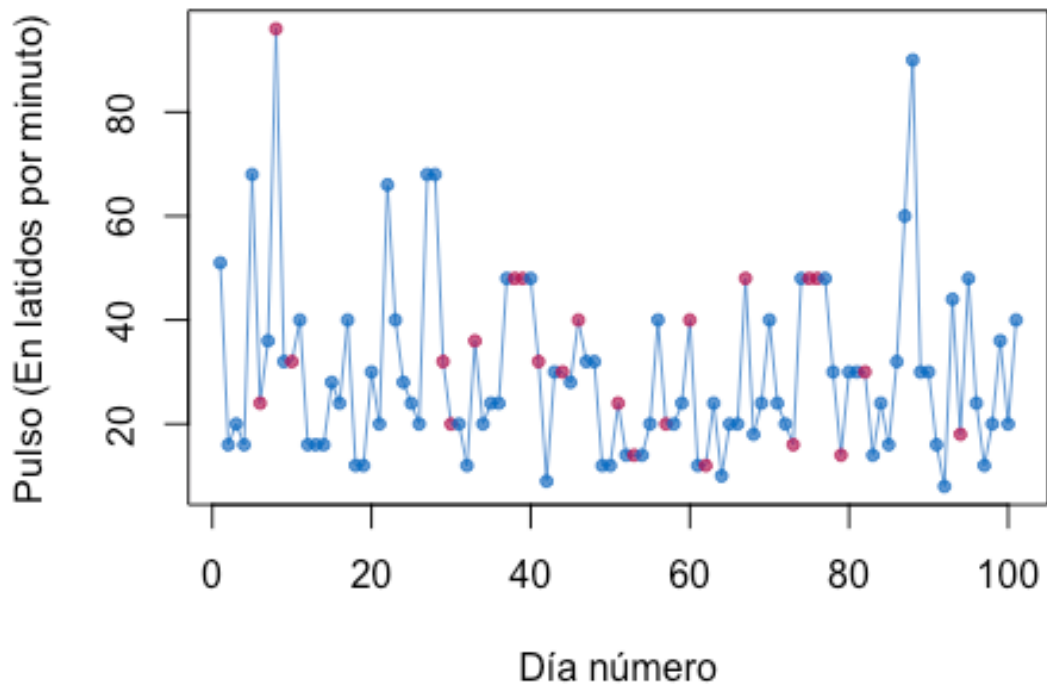


3.2.7 Imputación mediante LOCF y BOFC

```
# LOCF y BOFC
par(mfrow=c(1,1))
Pu <- subsample2$resp_rate
locf <- function(x) {
  a <- x[1]
  for (i in 2:length(x)) {
    if (is.na(x[i])) x[i] <- a
    else a <- x[i]
  }
  return(x)
}
Pul <- locf(Pu)
colvec <- ifelse(is.na(Pu),mdc(2),mdc(1))

# Ilustración 17
plot(Pul[100:200],col=colvec,type="l",xlab="Día número",ylab="Pulso (En
```

```
latidos por minuto)")
points(Pul[100:200],col=colvec,pch=20,cex=1)
```



3.2.8 Imputación múltiple

```
# Generación de subsample3 mediante paquete DPLYR
subsample3 <- slice(HOC,1:100) %>% select(pulse, resp_rate, pain,
packed_cell_volume, total_protein)
```

```
# Visualizo la estructura del dataset generado
str(subsample)
```

```
## 'data.frame':  10 obs. of  4 variables:
## $ pulse   : int  66 88 40 164 104 NA 48 60 80 90
## $ abd_dist: int  4 2 1 4 NA 2 3 2 4 1
## $ pain    : int  5 3 3 2 NA 2 3 NA 4 5
## $ nas_refl: int  NA NA NA 2 NA 1 1 1 1 1
```

```
# Observo el patrón de datos faltantes
```

```
pattern3 <- md.pattern(subsample3)
```

```
pattern3
```

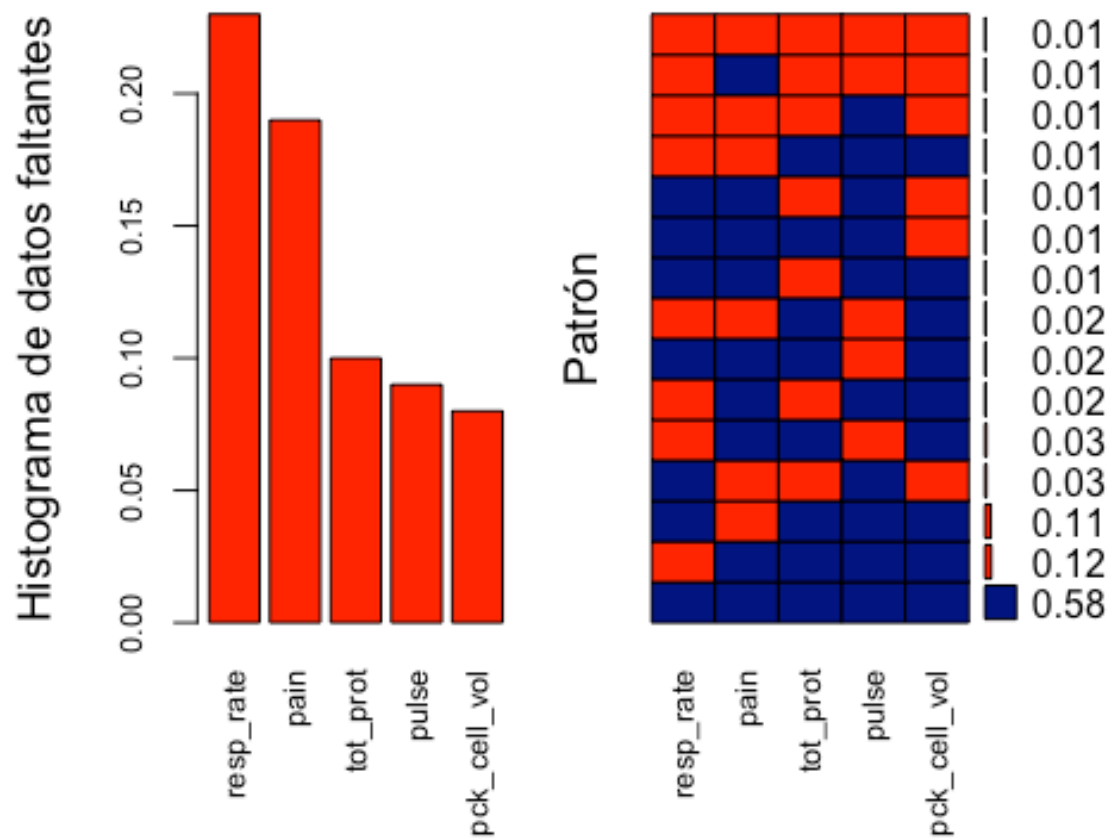
```
##      packed_cell_volume pulse total_protein pain resp_rate
## 58                    1      1              1    1         1  0
##  2                    1      0              1    1         1  1
## 12                    1      1              1    1         0  1
## 11                    1      1              1    0         1  1
##  1                    0      1              1    1         1  1
##  1                    1      1              0    1         1  1
##  3                    1      0              1    1         0  2
##  1                    1      1              1    0         0  2
##  2                    1      1              0    1         0  2
##  1                    0      1              0    1         1  2
##  2                    1      0              1    0         0  3
##  3                    0      1              0    0         1  3
##  1                    0      0              0    1         0  4
##  1                    0      1              0    0         0  4
##  1                    0      0              0    0         0  5
##                        8      9              10   19         23 69
```

```
# Abrevio los nombres de las variables de subsample3
```

```
names(subsample3) <- c("pulse", "resp_rate", "pain", "pck_cell_vol",  
"tot_prot")
```

```
# Ilustración 20: Visualización de patrón de valores faltantes en  
subsample3, mediante paquete VIM
```

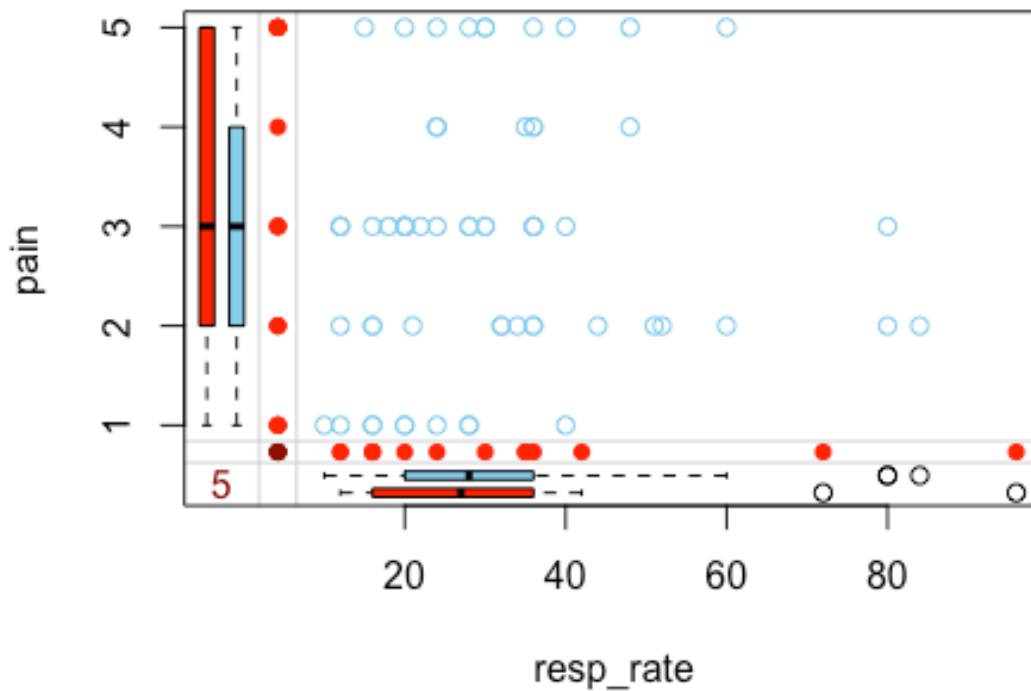
```
aggr_plot <- aggr(subsample3, col=c('navyblue', 'red'), numbers=TRUE,  
sortVars=TRUE, labels=names(subsample3), cex.axis=.8, gap=3,  
ylab=c("Histograma de datos faltantes", "Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable Count
## resp_rate 0.23
## pain 0.19
## tot_prot 0.10
## pulse 0.09
## pck_cell_vol 0.08
```

Ilustración 21: Visualización de relación entre valores faltantes de las variables resp_rate y pain del dataframe subsample3, mediante paquete DPLYR y VIM

```
subsample3 %>% select(resp_rate, pain) %>% marginplot()
```



```
# Imputación múltiple del dataframe subsample3
imp <- mice(subsample3, seed=1, print=FALSE)
fit <- with(imp, lm(pulse ~ resp_rate + pain + pck_cell_vol + tot_prot))
tab <- round(summary(pool(fit)),3)
tab[,c(1:3,5)]

##               est      se      t Pr(>|t|)
## (Intercept) -17.543 15.719 -1.116   0.276
## resp_rate    0.855  0.148  5.760   0.000
## pain         1.881  2.010  0.936   0.353
## pck_cell_vol  1.210  0.284  4.268   0.000
## tot_prot     0.153  0.104  1.463   0.153

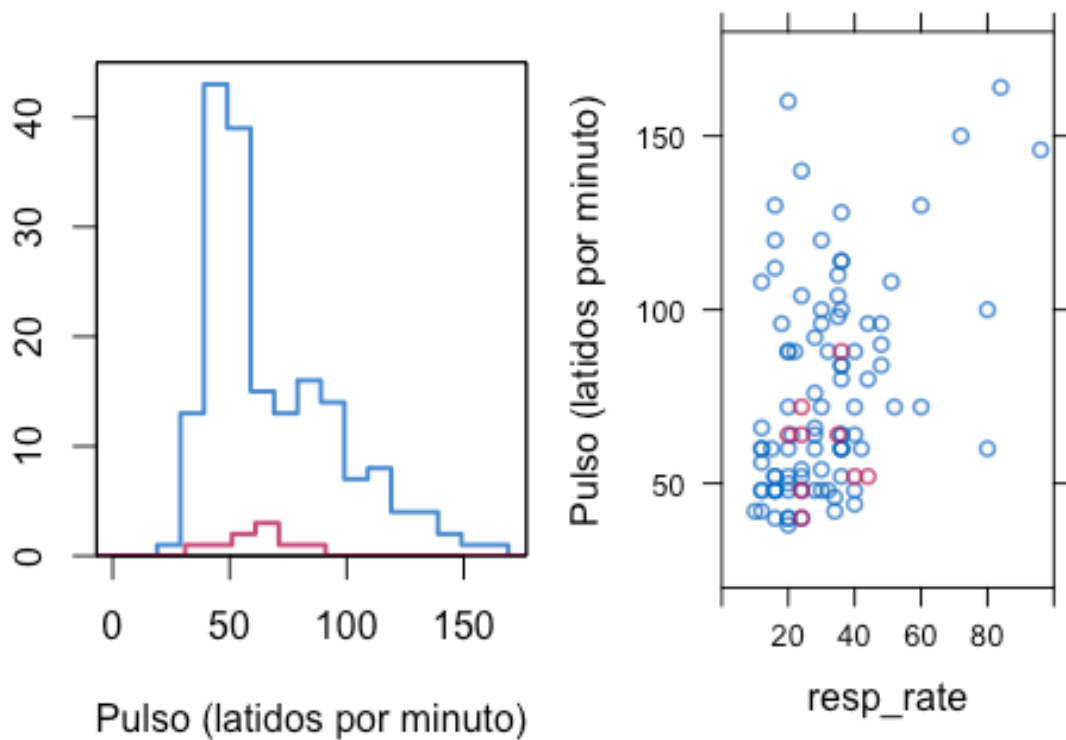
# Ilustración 22
par(mfrow = c(1,2))
mis <- imp$imp$pulse[,1]
fmis <- c(hist(mis, breaks, plot=FALSE)$counts, 0)
y <- matrix(c(fobs, fmis), ncol=2)
matplot(x, y, type="s",
        col=c(mdc(4),mdc(5)), lwd=2, lty=1,
        xlim = c(0, 170), ylim = c(0,45), yaxs = "i",
        xlab="Pulso (latidos por minuto)",
        ylab="Frecuencia")
```

```
box()
```

```
tp <- xyplot(imp, pulse ~ resp_rate, subset = .imp==1,  
  ylab="Pulso (latidos por minuto)", xlab="resp_rate",  
  cex = 0.75, lex=lwd,  
  ylim = c(20, 180), xlim = c(0,100))  
print(tp, newpage = FALSE, position = c(0.48,0.045,1,0.95))
```

```
# Verificación del modelo de imputación empleado en cada variable  
imp$meth
```

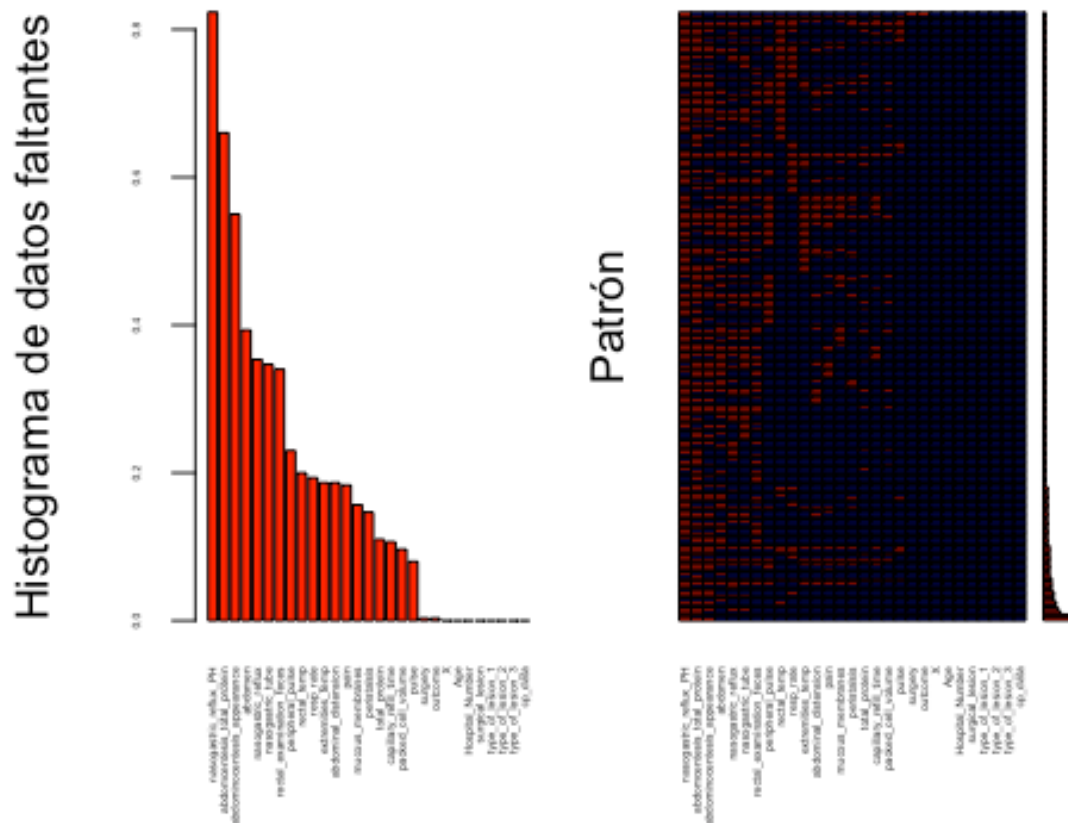
```
##      pulse      resp_rate      pain pck_cell_vol      tot_prot  
##      "pmm"      "pmm"      "pmm"      "pmm"      "pmm"
```



4 Imputación múltiple sobre los datasets abordados

4.1 Dataset "HOC"

```
# Dataset HOC (Horse-colic)
# Ilustración 23: Visualización de patrón de datos faltantes del dataset HOC
aggr_plot <- aggr(HOC, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(HOC), cex.axis=.3, gap=3, ylab=c("Histograma
de datos faltantes", "Patrón"))
```



```
##
## Variables sorted by number of missings:
##      Variable      Count
## nasogastric_reflux_PH 0.823333333
## abdomcentesis_total_protein 0.660000000
## abdominocentesis_appearance 0.550000000
##      abdomen 0.393333333
##      nasogastric_reflux 0.353333333
##      nasogastric_tube 0.346666667
##      rectal_examination_feces 0.340000000
```



```
##           peripheral_pulse 0.230000000
##           rectal_temp 0.200000000
##           resp_rate 0.193333333
##           extremities_temp 0.186666667
##           abdominal_distension 0.186666667
##           pain 0.183333333
##           mucous_membranes 0.156666667
##           peristalsis 0.146666667
##           total_protein 0.110000000
##           capillary_refill_time 0.106666667
##           packed_cell_volume 0.096666667
##           pulse 0.080000000
##           surgery 0.003333333
##           outcome 0.003333333
##           X 0.000000000
##           Age 0.000000000
##           Hospital_Number 0.000000000
##           surgical_lesion 0.000000000
##           type_of_lesion_1 0.000000000
##           type_of_lesion_2 0.000000000
##           type_of_lesion_3 0.000000000
##           cp_data 0.000000000
```

Imputación de la media en el dataset HOC

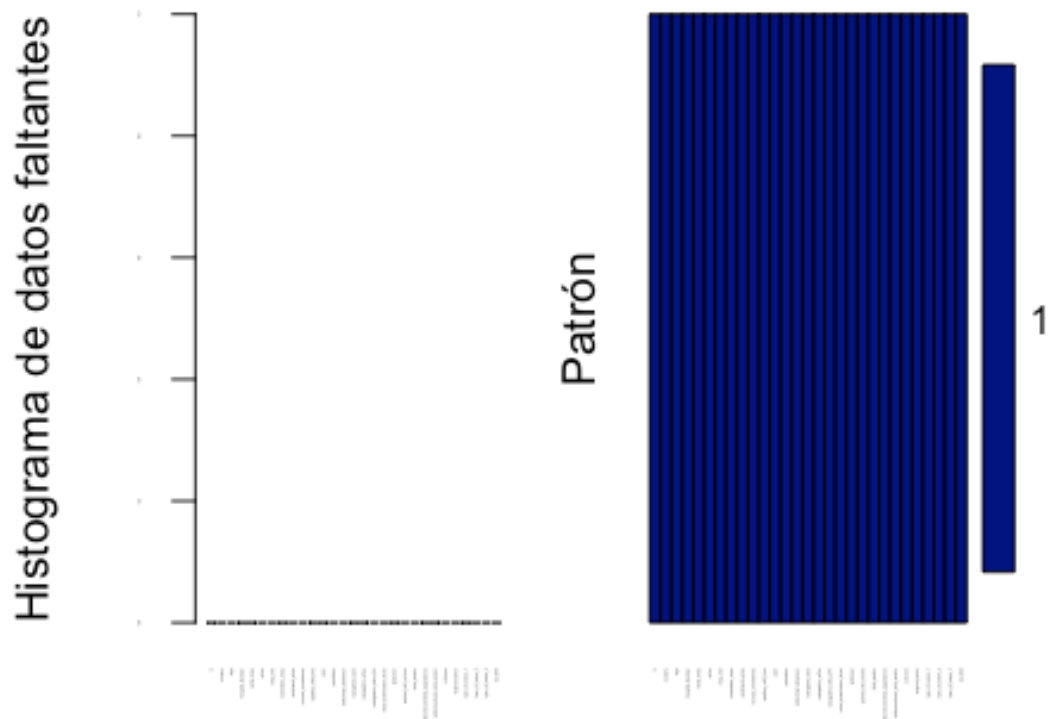
```
imp <- mice(HOC, method = "mean", seed=1, print=FALSE)
```

Imputación efectiva del dataset HOC

```
HOC_completed = complete(imp)
```

Ilustración 24: Visualización de patrón de datos en dataset HOC después de la imputación efectiva

```
aggr_plot <- aggr(HOC_completed, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(HOC_completed), cex.axis=.1, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
##      Variable Count
##      X           0
##      surgery     0
##      Age          0
##      Hospital_Number 0
##      rectal_temp  0
##      pulse        0
##      resp_rate    0
##      extremities_temp 0
##      peripheral_pulse 0
##      mucous_membranes 0
##      capillary_refill_time 0
##      pain         0
##      peristalsis  0
##      abdominal_distension 0
##      nasogastric_tube 0
##      nasogastric_reflux 0
##      nasogastric_reflux_PH 0
##      rectal_examination_feces 0
##      abdomen      0
##      packed_cell_volume 0
```

```
##          total_protein      0
## abdominocentesis_appearance  0
## abdomcentesis_total_protein  0
##          outcome          0
##          surgical_lesion     0
##          type_of_lesion_1     0
##          type_of_lesion_2     0
##          type_of_lesion_3     0
##          cp_data             0

# Verificación del modelo de imputación empleado en cada variable
imp$meth

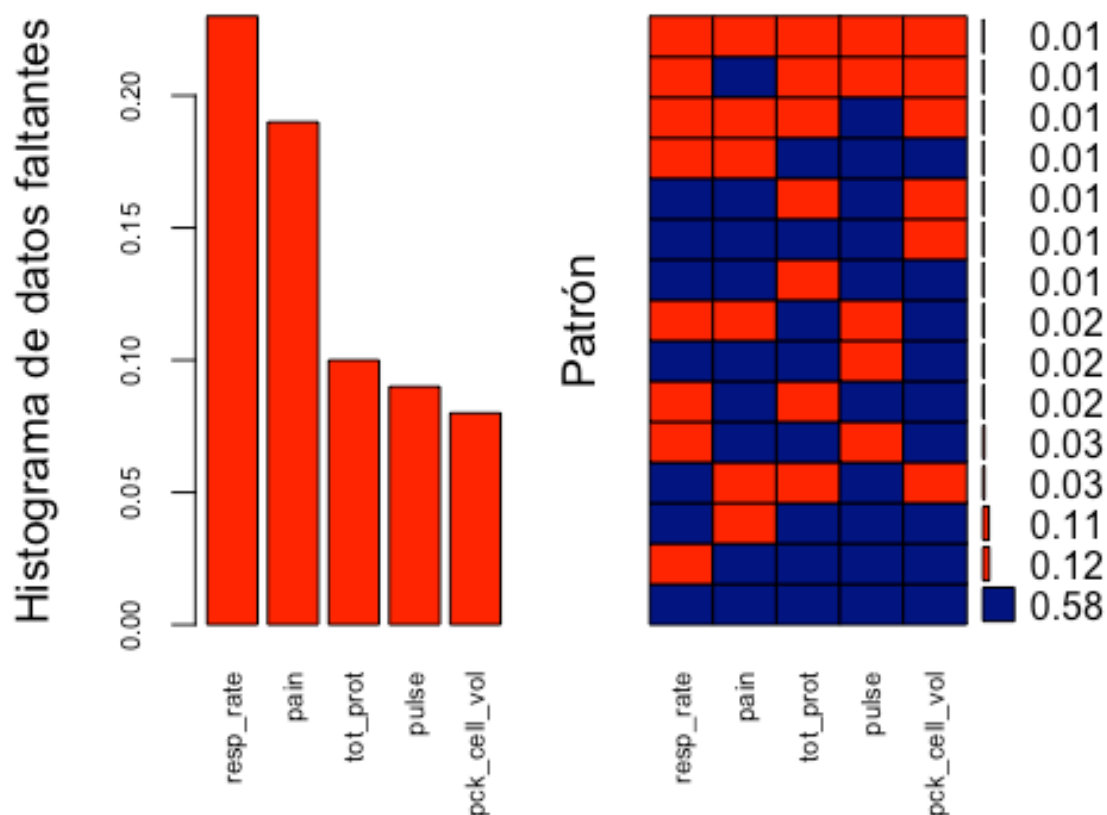
##          X                  surgery
##          "mean"              "mean"
##          Age                Hospital_Number
##          "mean"              "mean"
##          rectal_temp         pulse
##          "mean"              "mean"
##          resp_rate           extremities_temp
##          "mean"              "mean"
##          peripheral_pulse     mucous_membranes
##          "mean"              "mean"
##          capillary_refill_time pain
##          "mean"              "mean"
##          peristalsis         abdominal_distension
##          "mean"              "mean"
##          nasogastric_tube     nasogastric_reflux
##          "mean"              "mean"
##          nasogastric_reflux_PH rectal_examination_feces
##          "mean"              "mean"
##          abdomen            packed_cell_volume
##          "mean"              "mean"
##          total_protein abdominocentesis_appearance
##          "mean"              "mean"
## abdomcentesis_total_protein outcome
##          "mean"              "mean"
##          surgical_lesion     type_of_lesion_1
##          "mean"              "mean"
##          type_of_lesion_2     type_of_lesion_3
##          "mean"              "mean"
##          cp_data
##          "mean"
```

4.2 Dataset "subsample3"

```
# Dataset subsample3 (Muestra del dataset Horse-colic)
# Ilustración 23: Visualización de patrón de datos faltantes del dataset
```

HOC

```
aggr_plot <- aggr(subsample3, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(subsample3), cex.axis=.7, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable Count
## resp_rate 0.23
## pain 0.19
## tot_prot 0.10
## pulse 0.09
## pck_cell_vol 0.08
```

Imputación múltiple del dataset subsample3

```
imp <- mice(subsample3, method = "pmm", seed=1, print=FALSE)
```

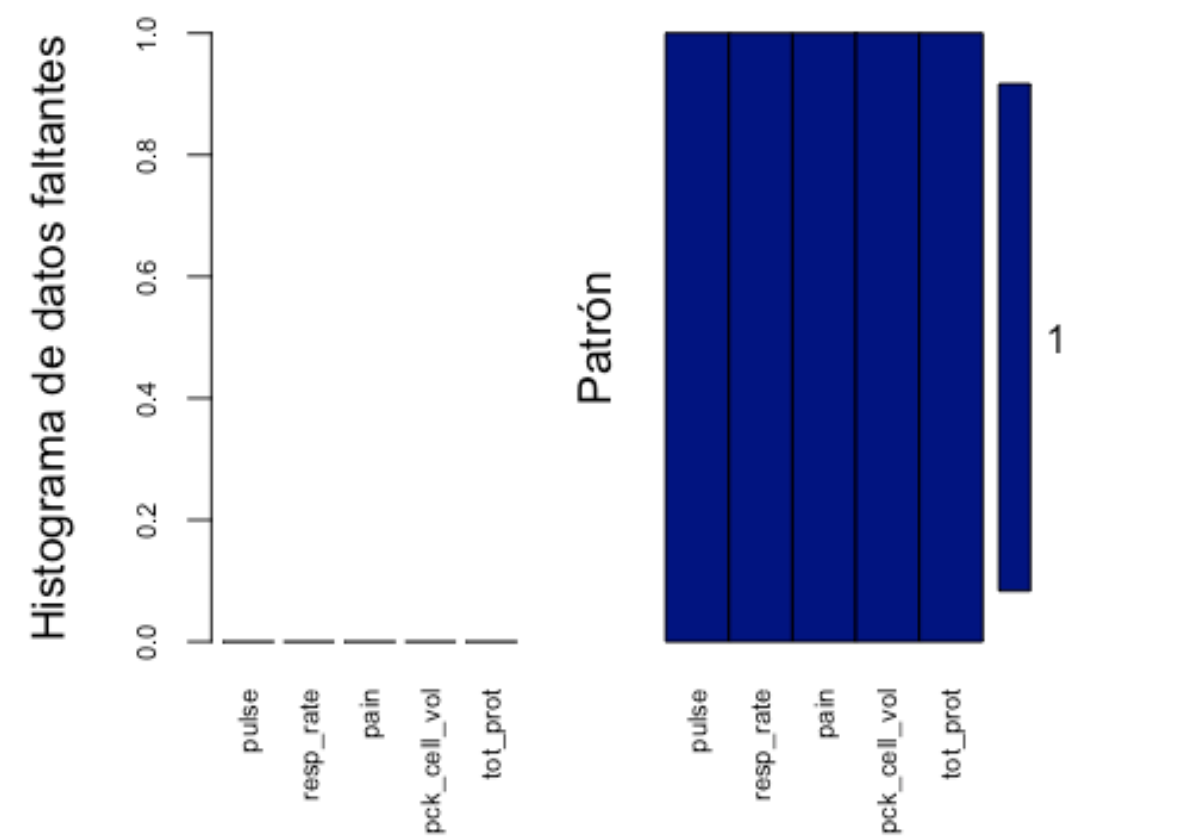
Imputación efectiva del dataset subsample3

```
subsample3_completed = complete(imp)
```

Ilustración XX: Visualización de patrón de datos en dataset subsample3 después de la imputación efectiva

```
aggr_plot <- aggr(subsample3_completed, col=c('navyblue','red'),
```

```
numbers=TRUE, sortVars=TRUE, labels=names(subsample3_completed),
cex.axis=.7, gap=3, ylab=c("Histograma de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable Count
## pulse 0
## resp_rate 0
## pain 0
## pck_cell_vol 0
## tot_prot 0

# Verificación del modelo de imputación empleado en cada variable
imp$meth

## pulse resp_rate pain pck_cell_vol tot_prot
## "pmm" "pmm" "pmm" "pmm" "pmm"
```

4.3 Dataset "AIR"

```
# Cargo librería datasets
library(datasets)

# Dataset AIR (Air Quality)
data(airquality)

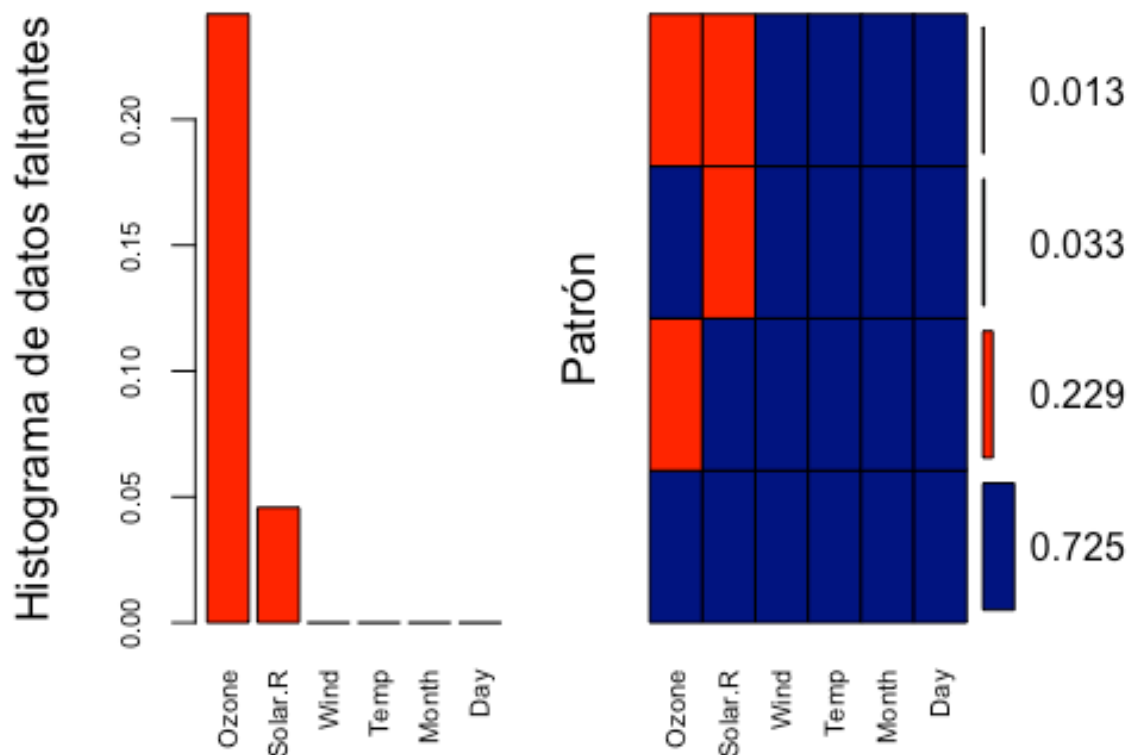
# Estructura del dataset
str(airquality)

## 'data.frame':    153 obs. of  6 variables:
## $ Ozone   : int  41 36 12 18 NA 28 23 19 8 NA ...
## $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...
## $ Wind    : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
## $ Temp    : int  67 72 74 62 56 66 65 59 61 69 ...
## $ Month   : int   5 5 5 5 5 5 5 5 5 5 ...
## $ Day     : int   1 2 3 4 5 6 7 8 9 10 ...

# Guardo el dataset AIR para generar el fichero AIR.rds
# (a efectos de reproducibilidad posterior)
saveRDS(airquality, "AIR.rds")

# Apertura del dataset AIR
# (Descomentar la línea siguiente si fuera necesario abrirlo para
# reproducibilidad posterior)
# airquality <- readRDS("AIR.rds")

# Ilustración 24: Visualización de patrón de datos faltantes del dataset
AIR
aggr_plot <- aggr(airquality, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(airquality), cex.axis=.7, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable      Count
## Ozone 0.24183007
## Solar.R 0.04575163
## Wind 0.00000000
## Temp 0.00000000
## Month 0.00000000
## Day 0.00000000

# Imputación múltiple del dataframe AIR
imp <- mice(airquality, seed=1, print=FALSE)
fit <- with(imp, lm(Ozone ~ Solar.R + Wind + Temp + Month + Day))
tab <- round(summary(pool(fit)), 3)
tab[, c(1:3, 5)]

##           est      se      t Pr(>|t|)
## (Intercept) -65.713 22.906 -2.869  0.006
## Solar.R      0.055  0.023  2.392  0.020
## Wind        -2.995  0.647 -4.630  0.000
## Temp         1.838  0.262  7.003  0.000
## Month       -2.757  1.704 -1.618  0.118
## Day          0.225  0.212  1.062  0.291
```

```

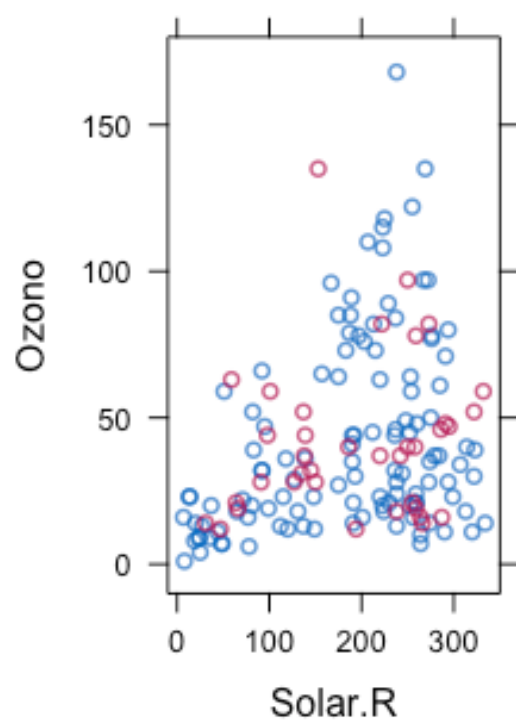
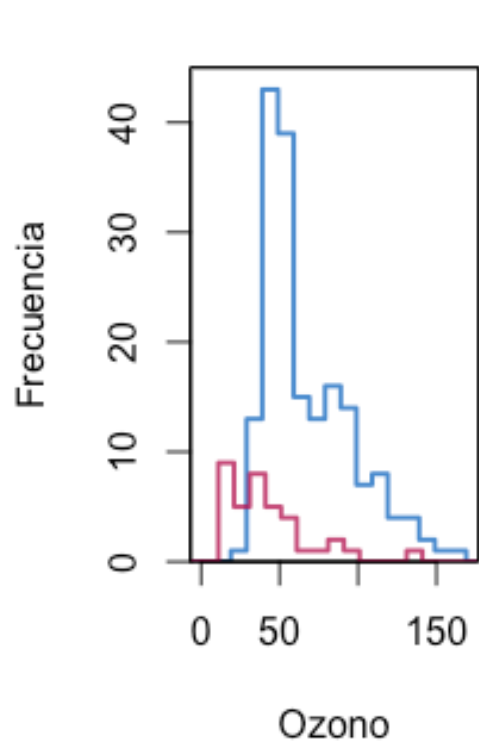
# Ilustración XX
par(mfrow = c(1,2))
mis <- imp$imp$Ozone[,1]
fmis <- c(hist(mis, breaks, plot=FALSE)$counts, 0)
y <- matrix(c(fobs, fmis), ncol=2)
matplot(x, y, type="s",
        col=c(mdc(4),mdc(5)), lwd=2, lty=1,
        xlim = c(0, 170), ylim = c(0,45), yaxs = "i",
        xlab="Ozono",
        ylab="Frecuencia")
box()

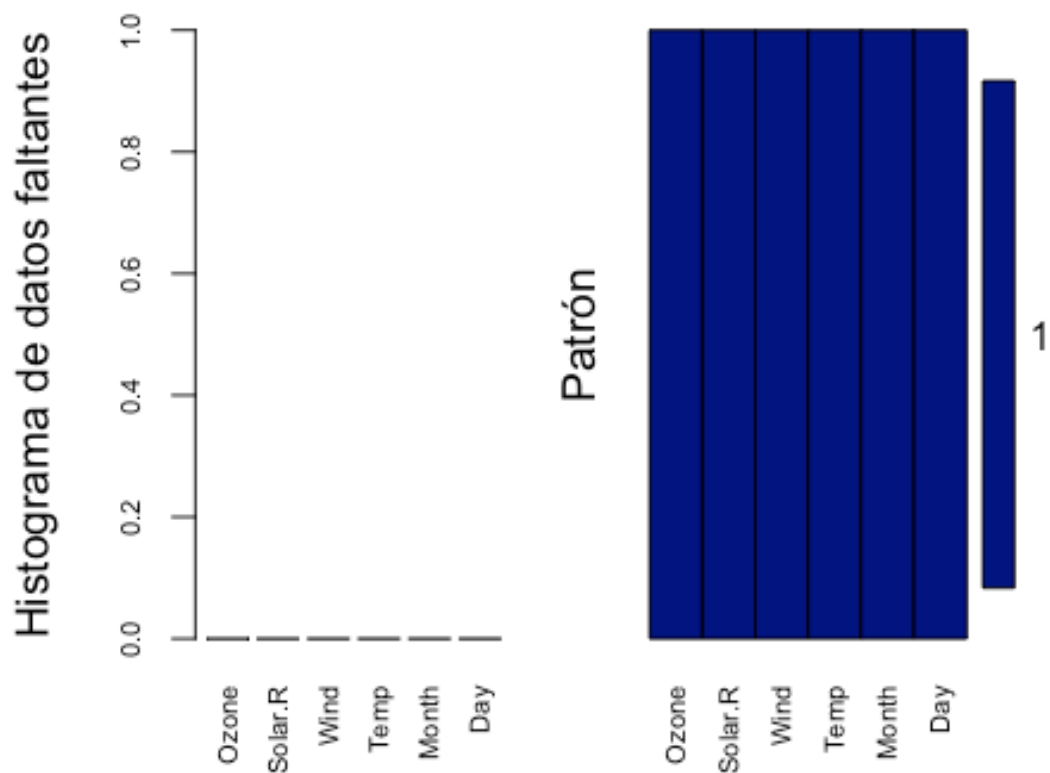
tp <- xyplot(imp, Ozone ~ Solar.R, subset = .imp==1,
            ylab="Ozono", xlab="Solar.R",
            cex = 0.75, lex=lwd,
            ylim = c(-10, 180), xlim = c(-10,350))
print(tp, newpage = FALSE, position = c(0.48,0.045,1,0.95))

# Imputación efectiva del dataset AIR
airquality_completed = complete(imp)

# Ilustración XX: Visualización de patrón de datos en dataset AIR después
de la imputación efectiva
aggr_plot <- aggr(airquality_completed, col=c('navyblue','red'),
                numbers=TRUE, sortVars=TRUE, labels=names(airquality_completed),
                cex.axis=.7, gap=3, ylab=c("Histograma de datos faltantes","Patrón"))

```



```
##
## Variables sorted by number of missings:
## Variable Count
## Ozone      0
## Solar.R    0
## Wind       0
## Temp       0
## Month      0
## Day        0

# Verificación del modelo de imputación empleado en cada variable
imp$meth

## Ozone Solar.R Wind Temp Month Day
## "pmm" "pmm" "" "" "" ""

# Comparación de dataset original con dataset imputado
head(airquality)

## Ozone Solar.R Wind Temp Month Day
## 1 41 190 7.4 67 5 1
## 2 36 118 8.0 72 5 2
## 3 12 149 12.6 74 5 3
## 4 18 313 11.5 62 5 4
```

```
## 5      NA      NA 14.3   56     5   5
## 6      28      NA 14.9   66     5   6
```

```
head(airquality_completed) # Imputado con Regresión Lineal
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190   7.4   67     5   1
## 2    36     118   8.0   72     5   2
## 3    12     149  12.6   74     5   3
## 4    18     313  11.5   62     5   4
## 5    19     259  14.3   56     5   5
## 6    28     294  14.9   66     5   6
```

```
# Dataframe con Las 2 variables Ozone de Los 2 datasets
```

```
comp_Ozone <- data.frame(airquality$Ozone, airquality_completed$Ozone)
comp_Ozone <- melt(comp_Ozone)
```

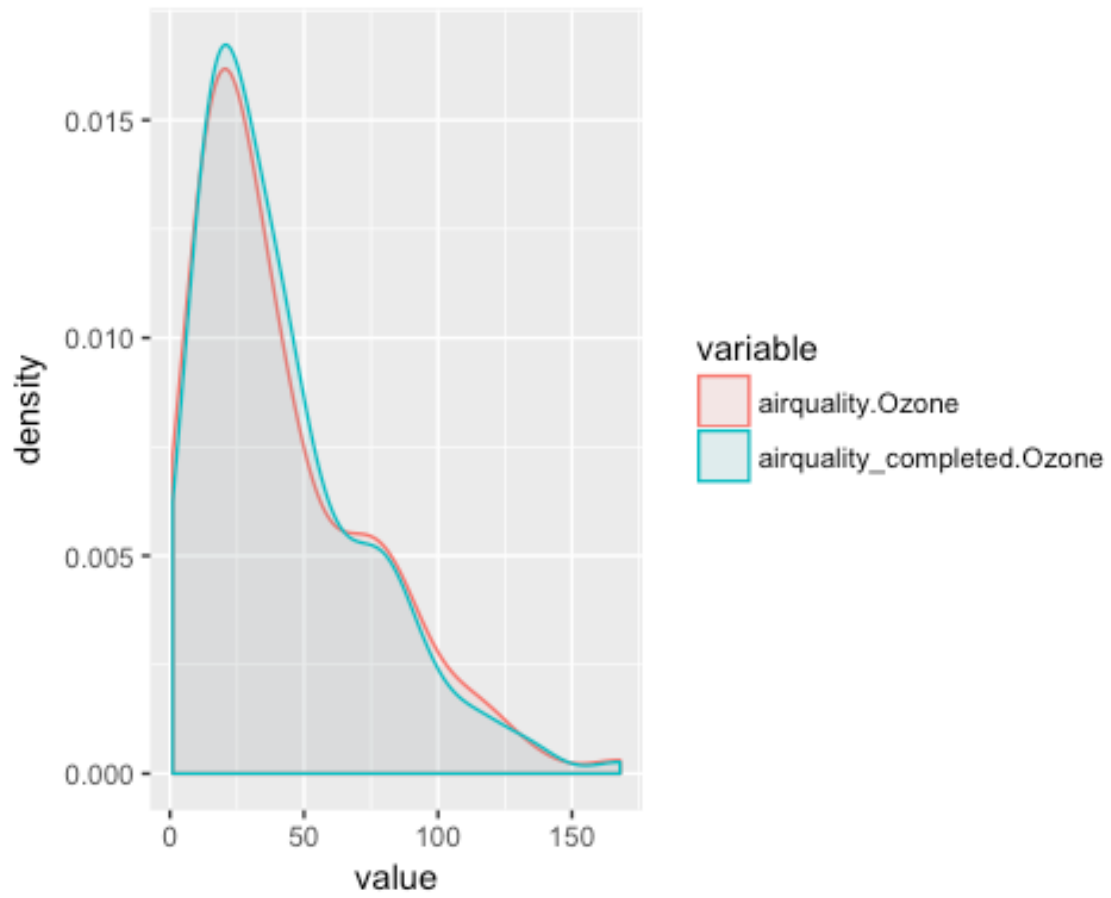
```
## No id variables; using all as measure variables
```

```
# Cargo Librería GGLOT2
```

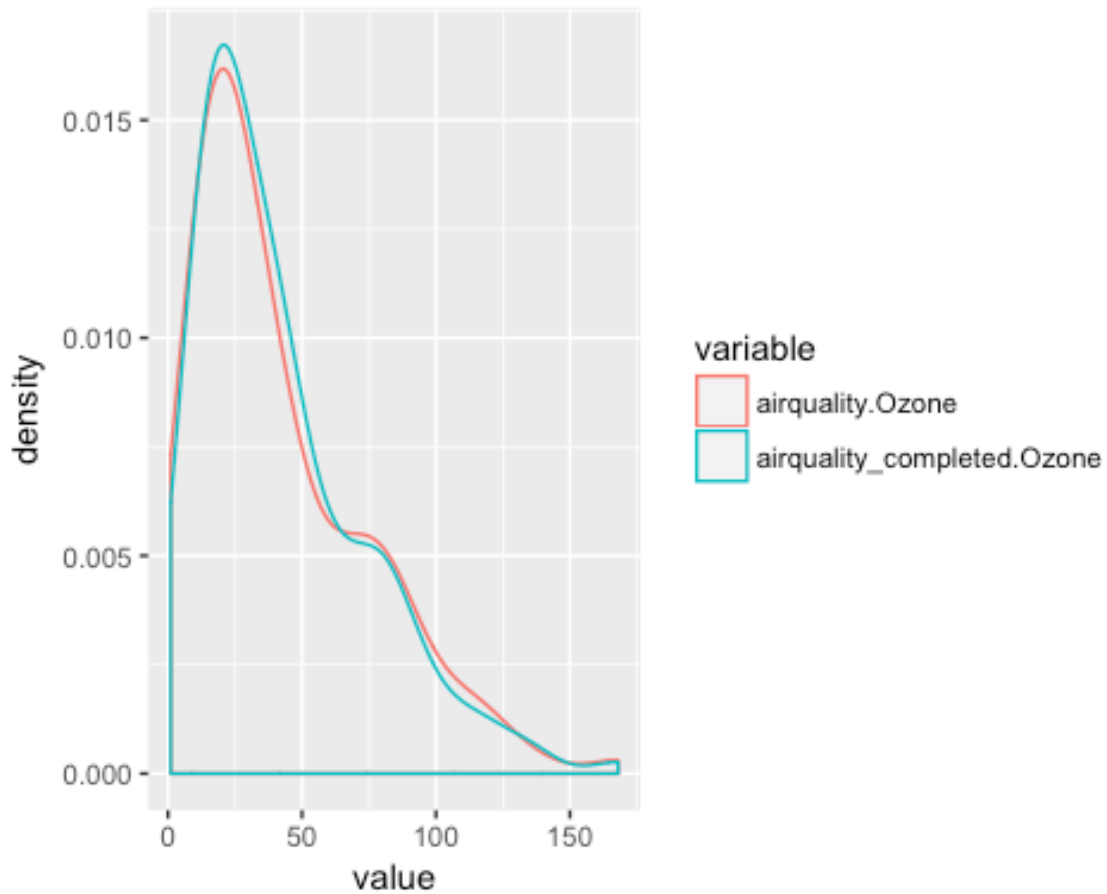
```
library(ggplot2)
```

```
# Plot de densidad con Las 2 variables Ozone de Los 2 datasets
```

```
ggplot(comp_Ozone , aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)
```



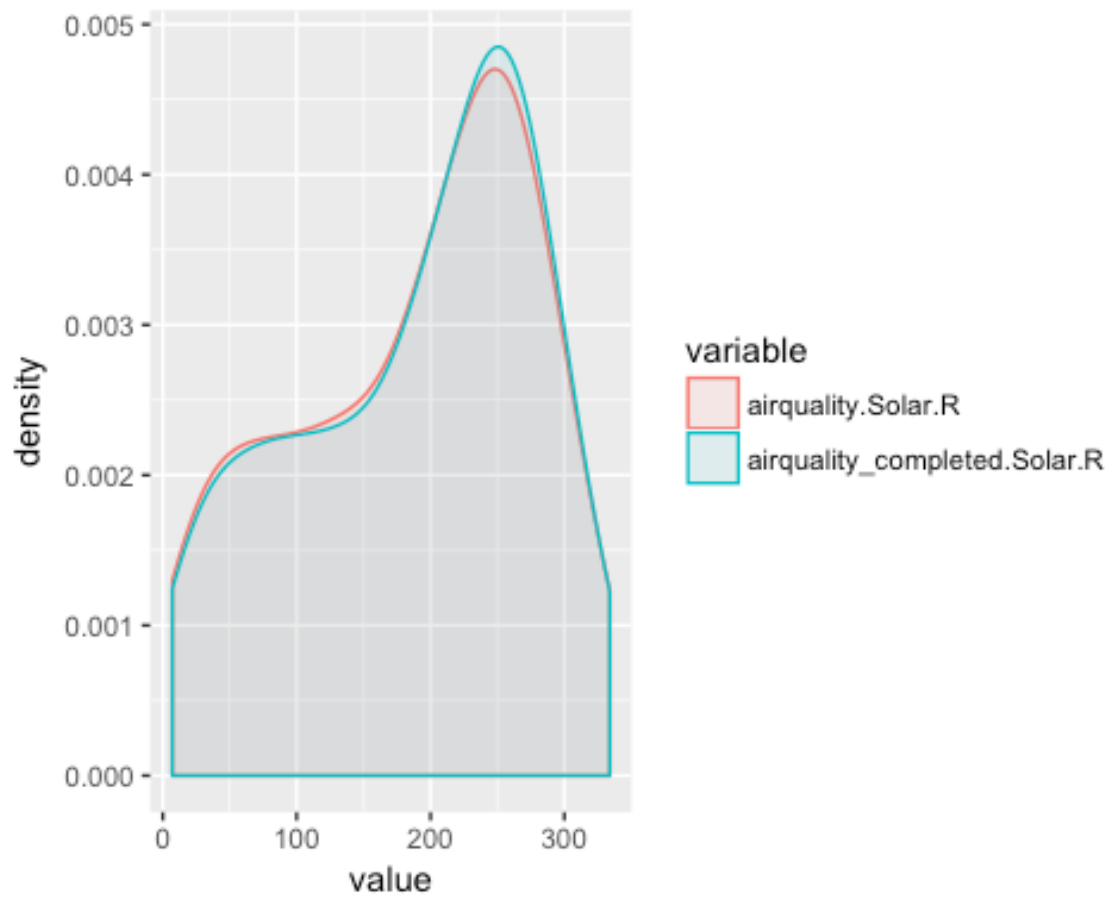
```
ggplot(comp_Ozone , aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



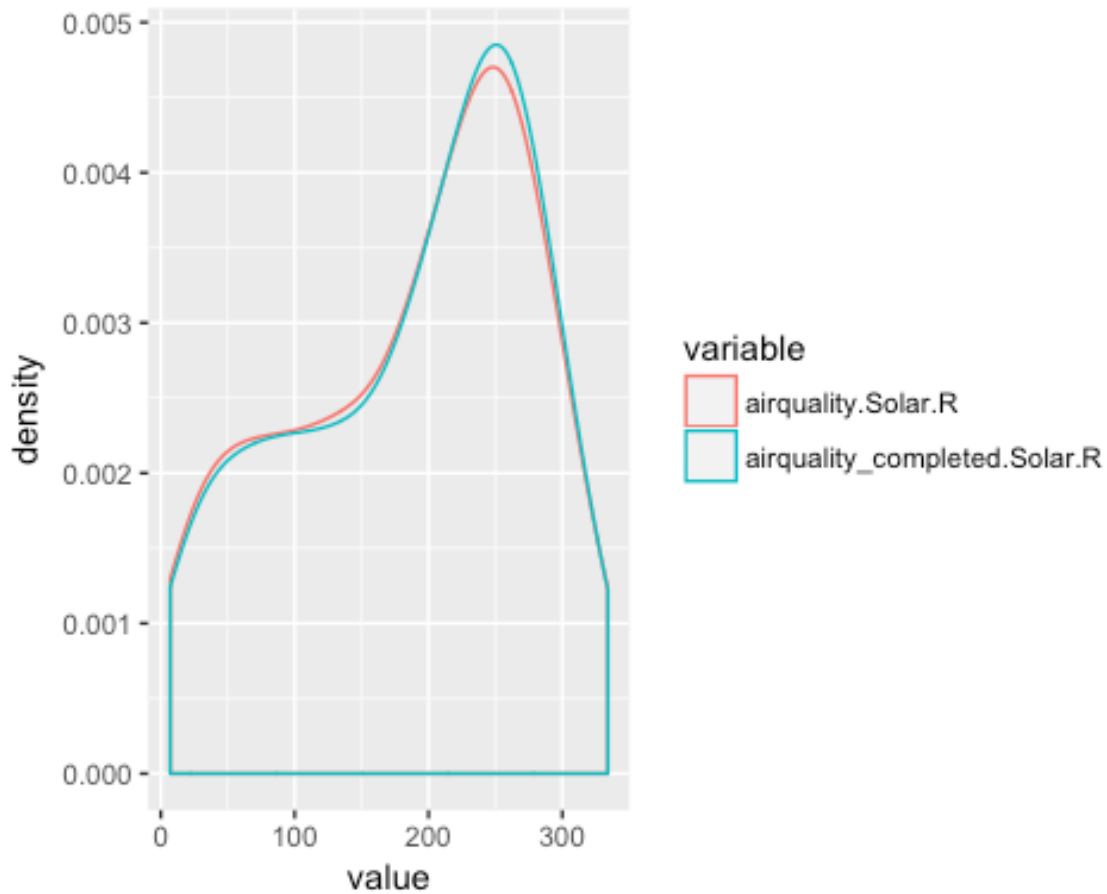
```
# Dataframe con Las 2 variables Solar.R de Los 2 datasets
comp_Solar.R <- data.frame(airquality$Solar.R,
airquality_completed$Solar.R)
comp_Solar.R <- melt(comp_Solar.R)

## No id variables; using all as measure variables

# Plot de densidad con Las 2 variables Solar.R de Los 2 datasets
ggplot(comp_Solar.R , aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)
```



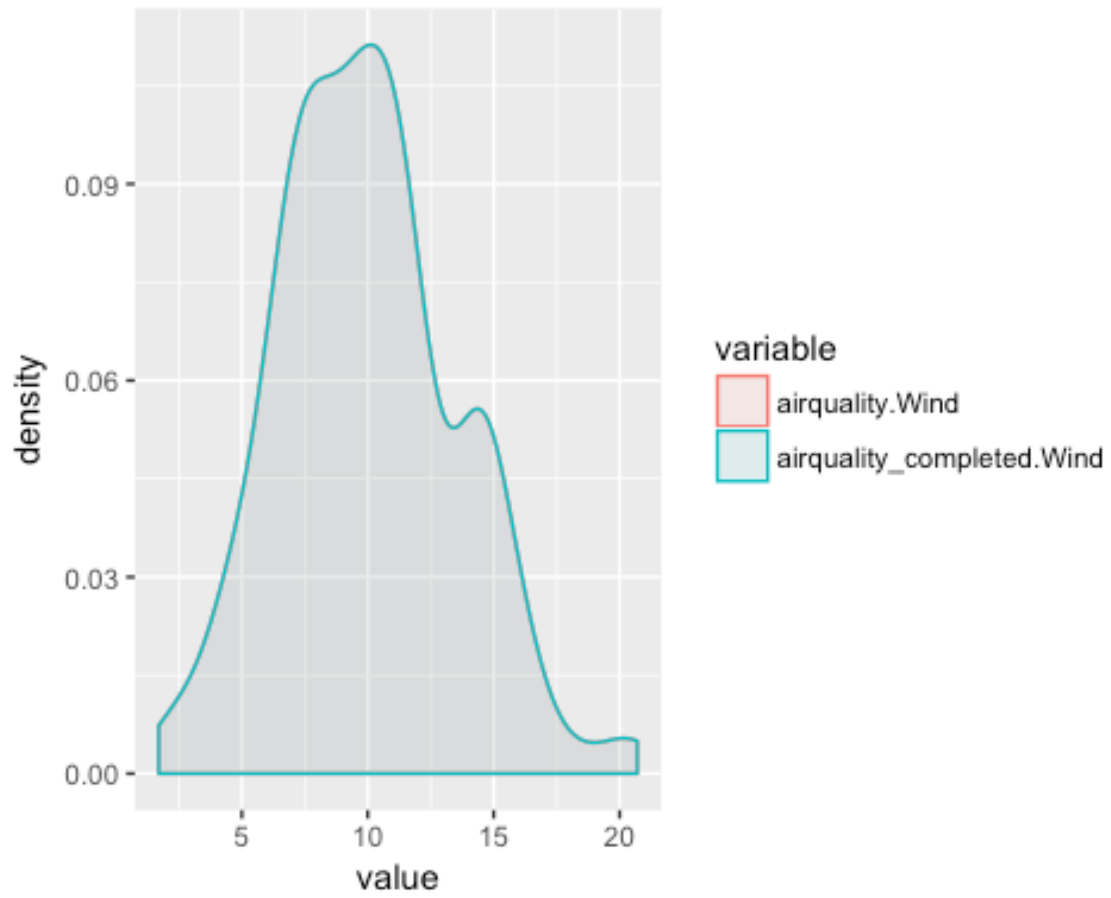
```
ggplot(comp_Solar.R , aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



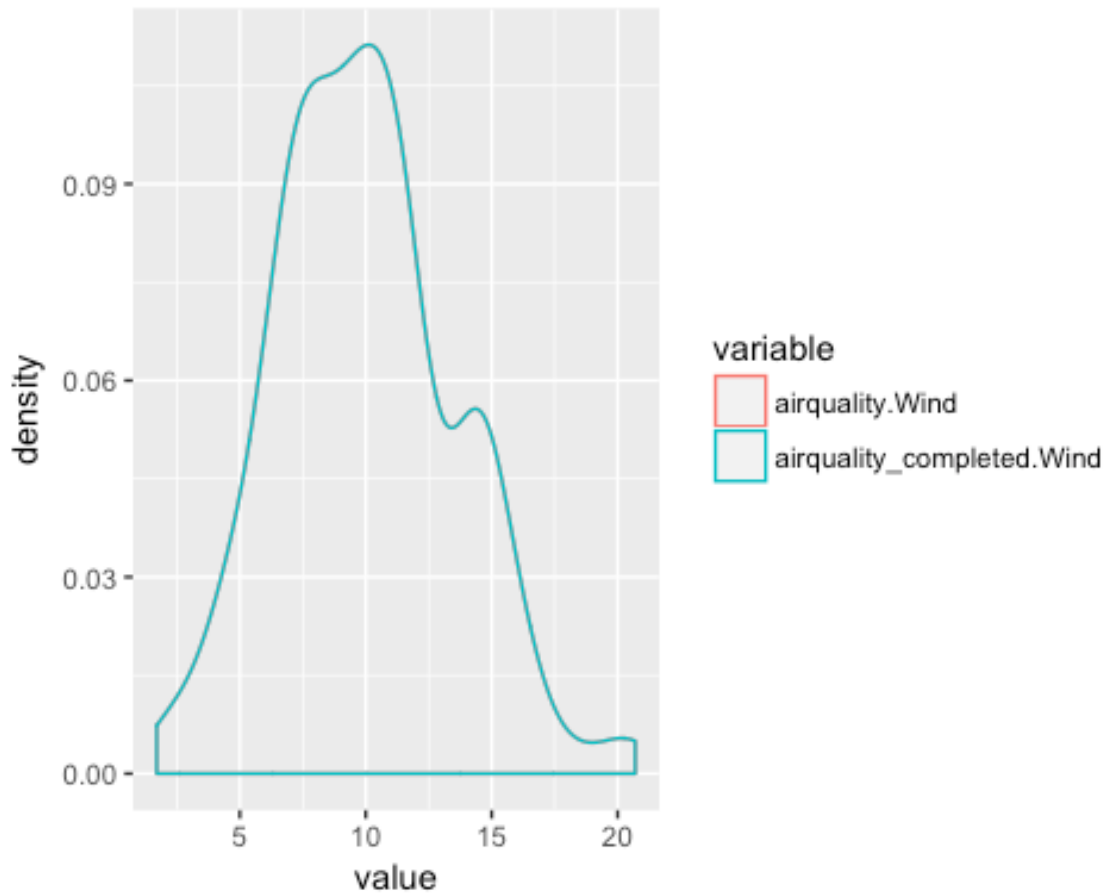
```
# Dataframe con Las 2 variables Wind de Los 2 datasets
comp_Wind <- data.frame(airquality$Wind, airquality_completed$Wind)
comp_Wind <- melt(comp_Wind)

## No id variables; using all as measure variables

# Plot de densidad con Las 2 variables Wind de Los 2 datasets (no se
# produjeron imputaciones porque Wind no tenía NAs)
ggplot(comp_Wind , aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)
```



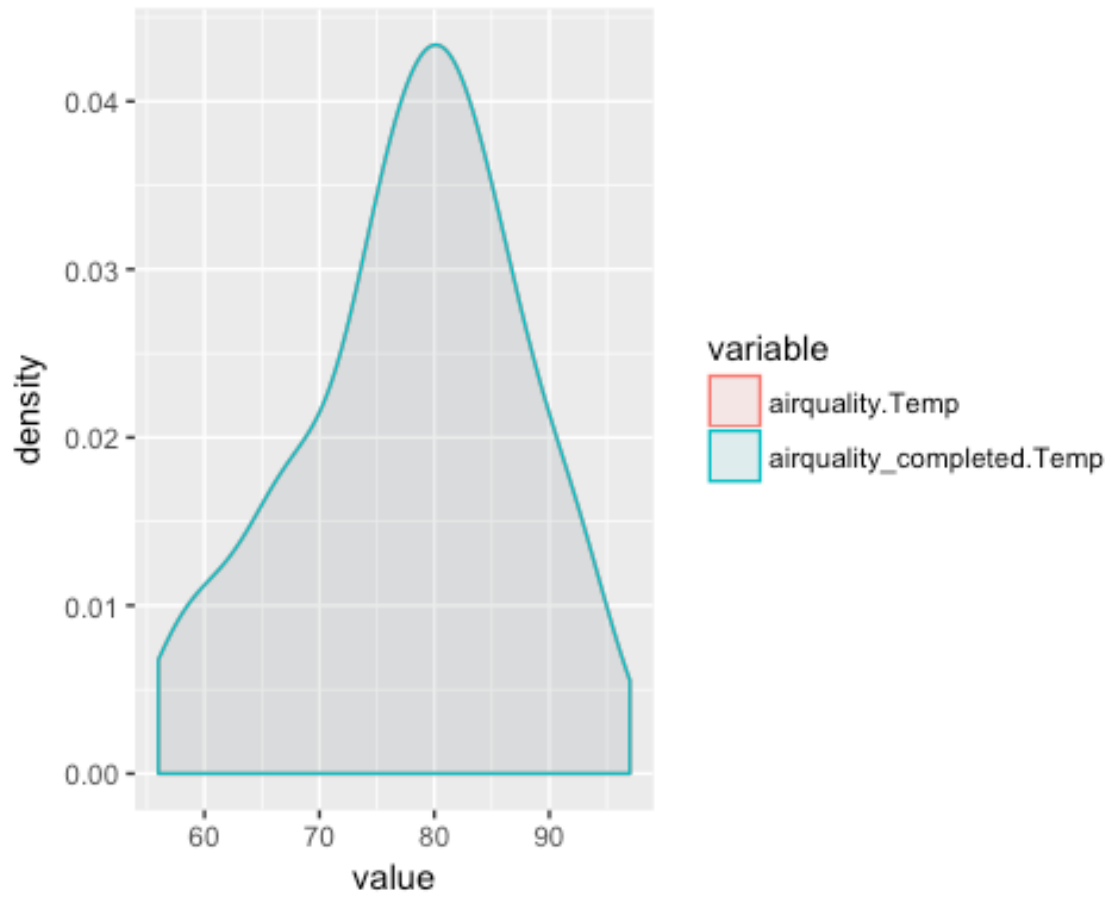
```
ggplot(comp_Wind , aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```

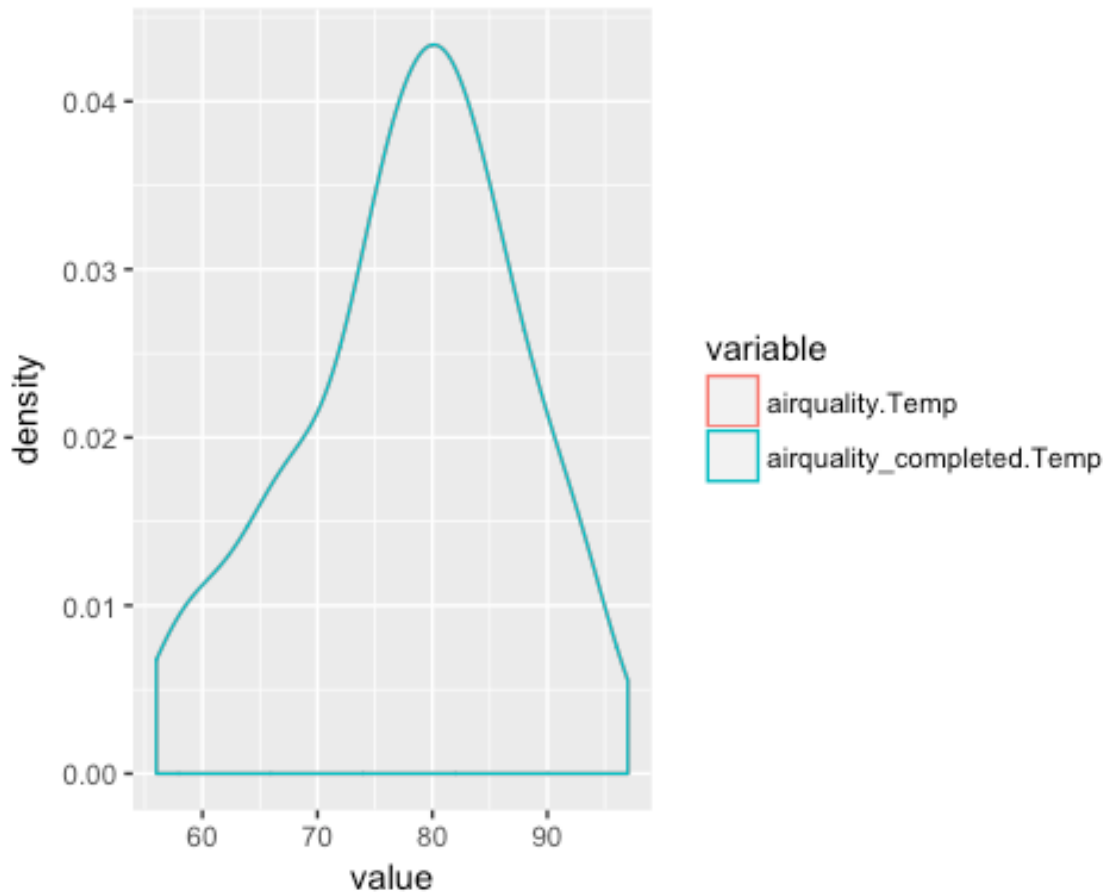
```
# Dataframe con Las 2 variables Temp de Los 2 datasets
comp_Temp <- data.frame(airquality$Temp, airquality_completed$Temp)
comp_Temp <- melt(comp_Temp)

## No id variables; using all as measure variables

# Plot de densidad con Las 2 variables Temp de Los 2 datasets (no se
# produjeron imputaciones porque Wind no tenía NAs)
ggplot(comp_Temp , aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)
```



```
ggplot(comp_Temp , aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



4.4 Dataset "NHAN"

Dataset NHAN (Nhanes)

data(nhanes)

Estructura del dataset

str(nhanes)

```
## 'data.frame':  25 obs. of  4 variables:
```

```
## $ age: num  1 2 1 3 1 3 1 1 2 2 ...
```

```
## $ bmi: num  NA 22.7 NA NA 20.4 NA 22.5 30.1 22 NA ...
```

```
## $ hyp: num  NA 1 1 NA 1 NA 1 1 1 NA ...
```

```
## $ chl: num  NA 187 187 NA 113 184 118 187 238 NA ...
```

Guardo el dataset NHAN para generar el fichero NHAN.rds

(a efectos de reproducibilidad posterior)

saveRDS(nhanes, "NHAN.rds")

Apertura del dataset NHAN

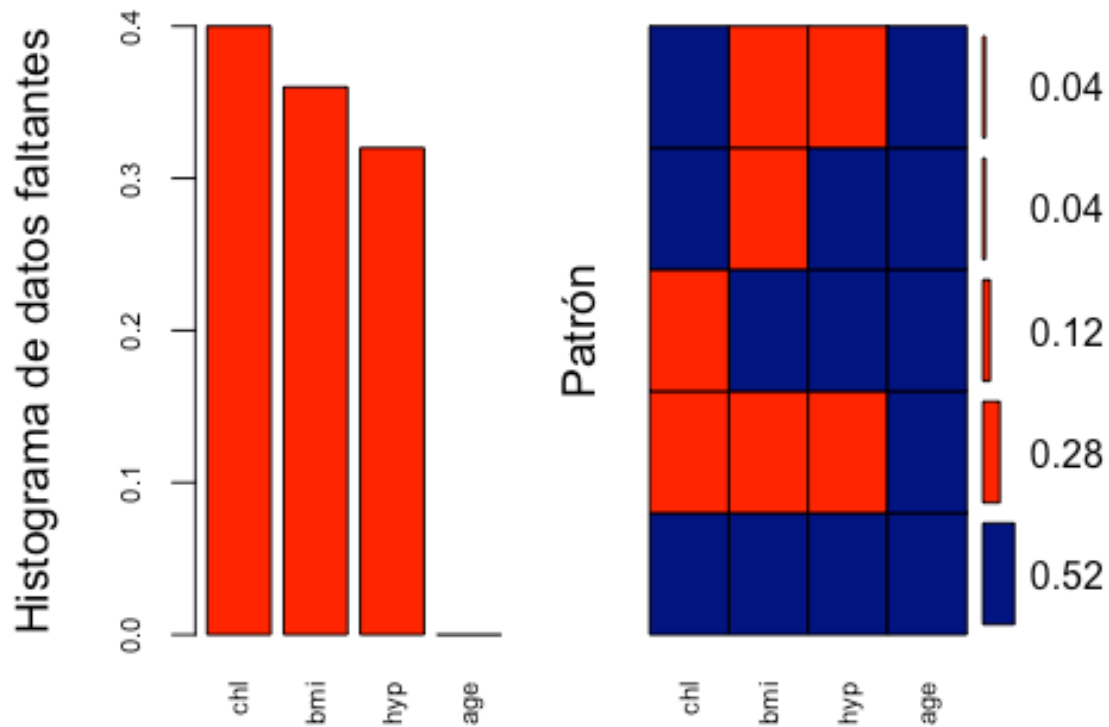
(Descomentar la línea siguiente si fuera necesario abrirlo para

```

reproducibilidad posterior)
# nhanes <- readRDS("NHAN.rds")

# Ilustración 27: Visualización de patrón de datos faltantes del dataset
NHAN
aggr_plot <- aggr(nhanes, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(nhanes), cex.axis=.7, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))

```



```

##
## Variables sorted by number of missings:
## Variable Count
##   chl  0.40
##   bmi  0.36
##   hyp  0.32
##   age  0.00

# Imputación múltiple del dataframe NHAN
imp <- mice(nhanes, seed=5, print=FALSE)
fit <- with(imp, lm(bmi ~ age + hyp + chl))
tab <- round(summary(pool(fit)),3)
tab[,c(1:3,5)]

```

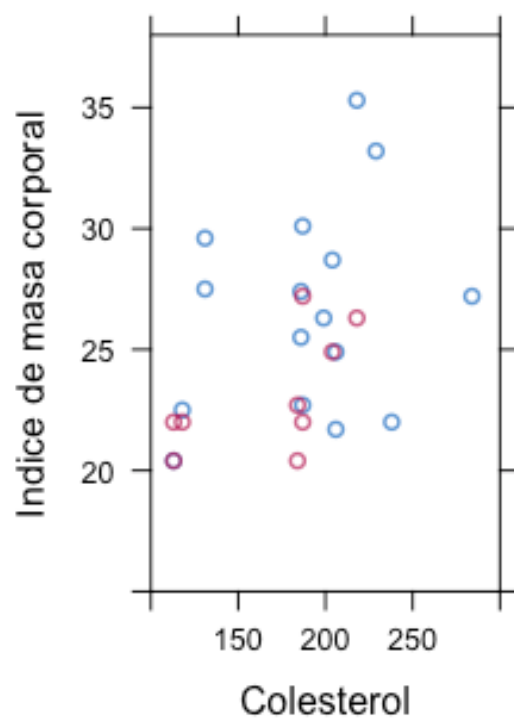
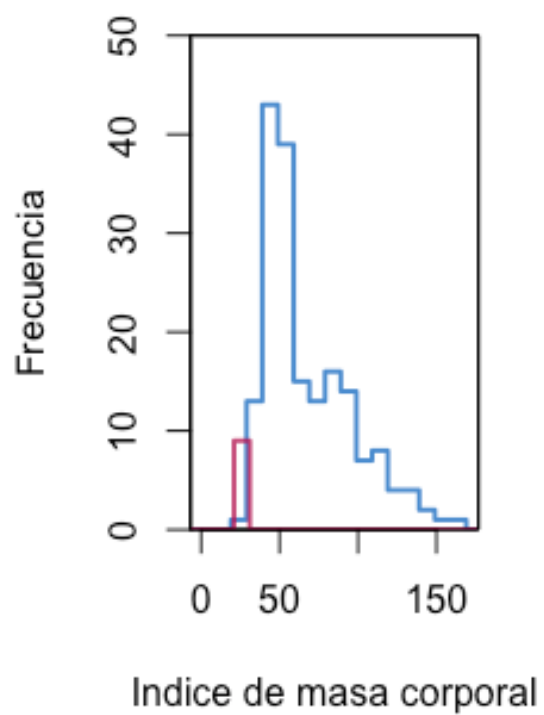
```
##               est      se      t Pr(>|t|)
## (Intercept) 20.594 3.983  5.170  0.002
## age         -3.784 1.318 -2.872  0.024
## hyp          1.604 1.876  0.855  0.408
## chl          0.054 0.022  2.494  0.040

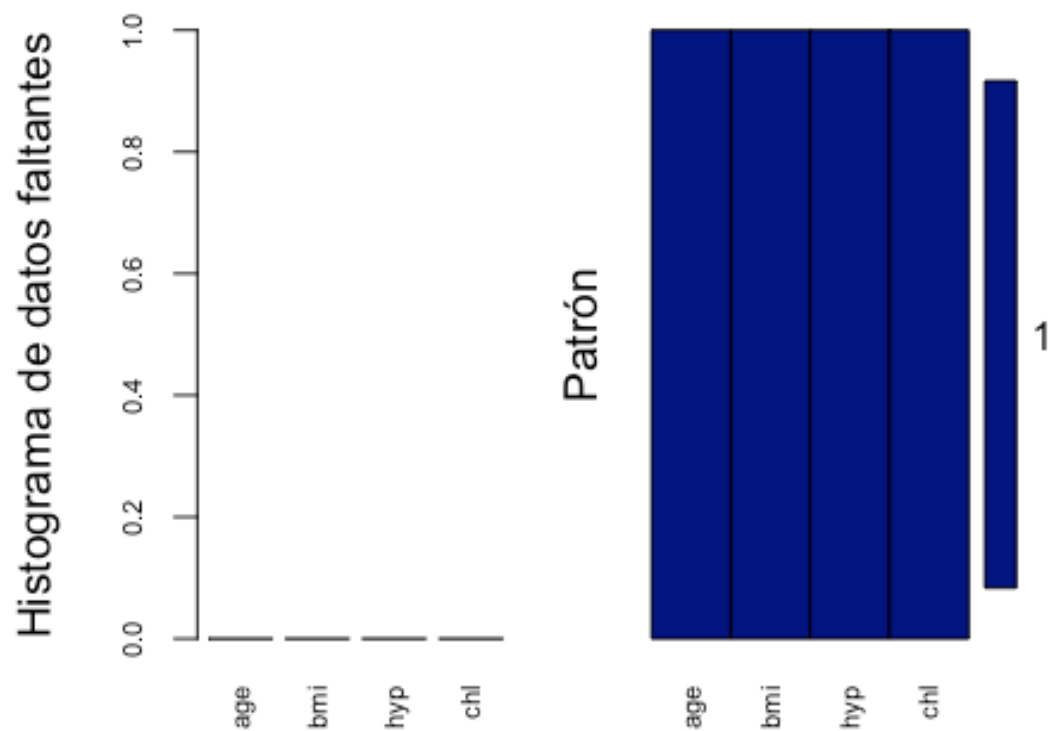
# Ilustración XX
par(mfrow = c(1,2))
mis <- imp$imp$bmi[,1]
fmis <- c(hist(mis, breaks, plot=FALSE)$counts, 0)
y <- matrix(c(fobs, fmis), ncol=2)
matplot(x, y, type="s",
        col=c(mdc(4),mdc(5)), lwd=2, lty=1,
        xlim = c(0, 170), ylim = c(0,50), yaxs = "i",
        xlab="Indice de masa corporal",
        ylab="Frecuencia")
box()

tp <- xyplot(imp, bmi ~ chl, subset = .imp==1,
            ylab="Indice de masa corporal", xlab="Colesterol",
            cex = 0.75, lex=lwd,
            ylim = c(15, 38), xlim = c(100,300))
print(tp, newpage = FALSE, position = c(0.48,0.045,1,0.95))

# Imputación efectiva del dataset NHAN
nhanes_completed = complete(imp)

# Ilustración XX: Visualización de patrón de datos en dataset NHAN
después de la imputación efectiva
aggr_plot <- aggr(nhanes_completed, col=c('navyblue','red'),
numbers=TRUE, sortVars=TRUE, labels=names(nhanes_completed), cex.axis=.7,
gap=3, ylab=c("Histograma de datos faltantes","Patrón"))
```





```
##
## Variables sorted by number of missings:
## Variable Count
##    age      0
##    bmi      0
##    hyp      0
##    chl      0

# Verificación del modelo de imputación empleado en cada variable
imp$meth

##    age    bmi    hyp    chl
##    "" "pmm" "pmm" "pmm"

# Comparación de dataset original con dataset imputado
head(nhanes)

##    age    bmi    hyp    chl
## 1     1     NA     NA     NA
## 2     2  22.7     1    187
## 3     1     NA     1    187
## 4     3     NA     NA     NA
## 5     1  20.4     1    113
## 6     3     NA     NA    184
```

```

head(nhanes_completed) # Imputado con PMM

##   age  bmi hyp chl
## 1   1 22.0   1 113
## 2   2 22.7   1 187
## 3   1 27.2   1 187
## 4   3 24.9   2 204
## 5   1 20.4   1 113
## 6   3 20.4   2 184

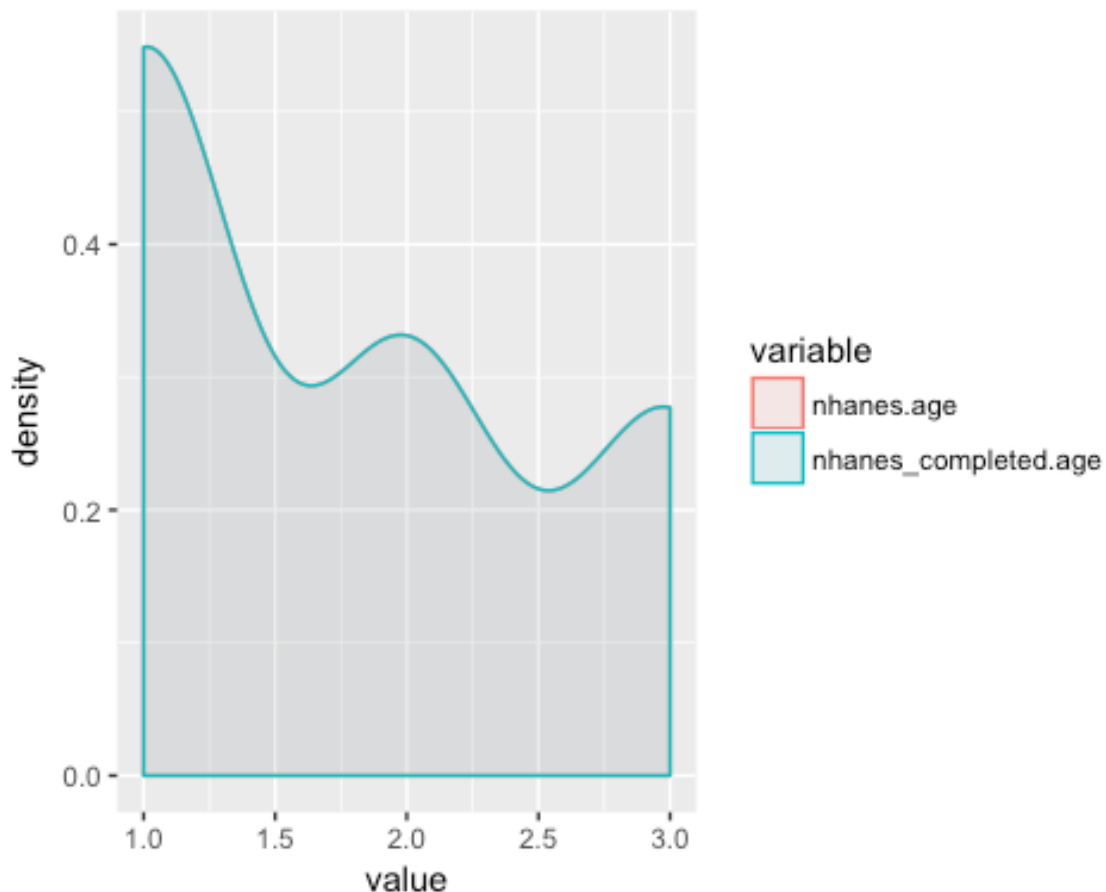
# Dataframe con las 2 variables age de Los 2 datasets
comp_age <- data.frame(nhanes$age, nhanes_completed$age)
comp_age <- melt(comp_age)

## No id variables; using all as measure variables

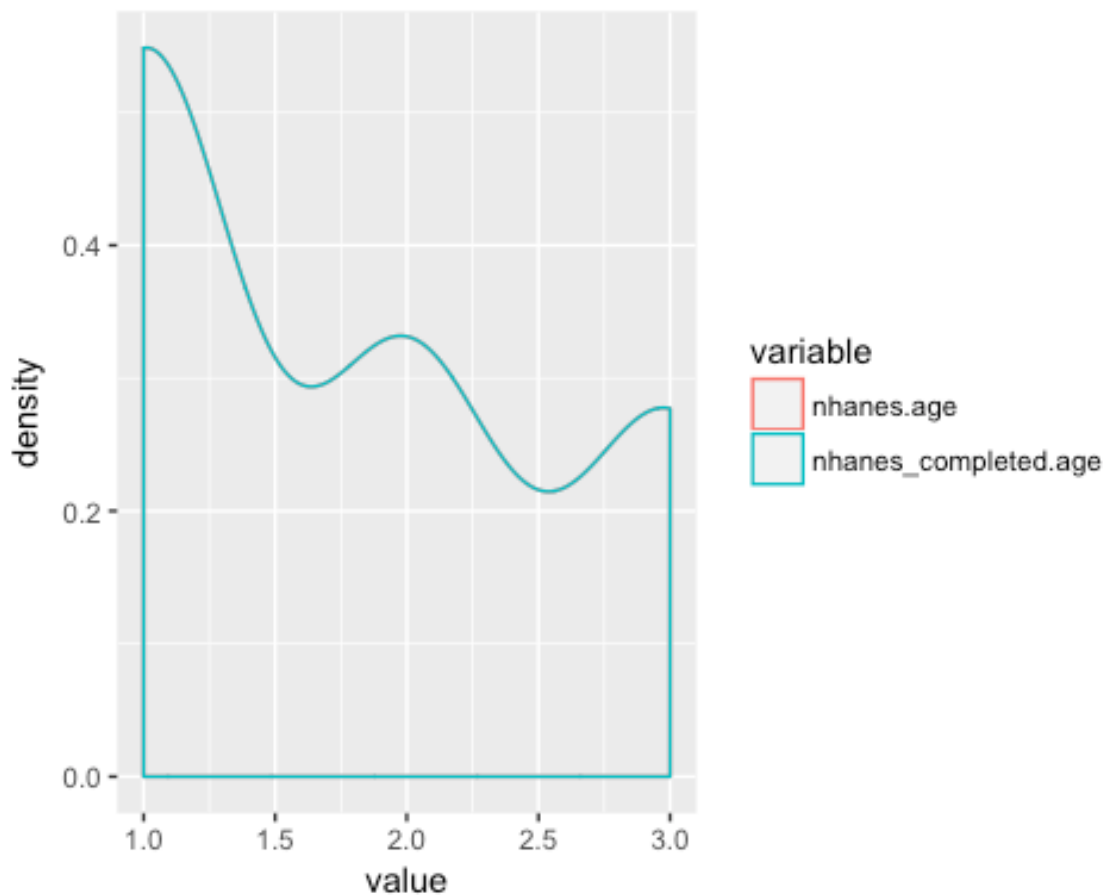
# Cargo librería GGLOT2
library(ggplot2)

# Plot de densidad con las 2 variables age de Los 2 datasets (no se
# produjeron imputaciones porque hyp no tenía NAs)
ggplot(comp_age, aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)

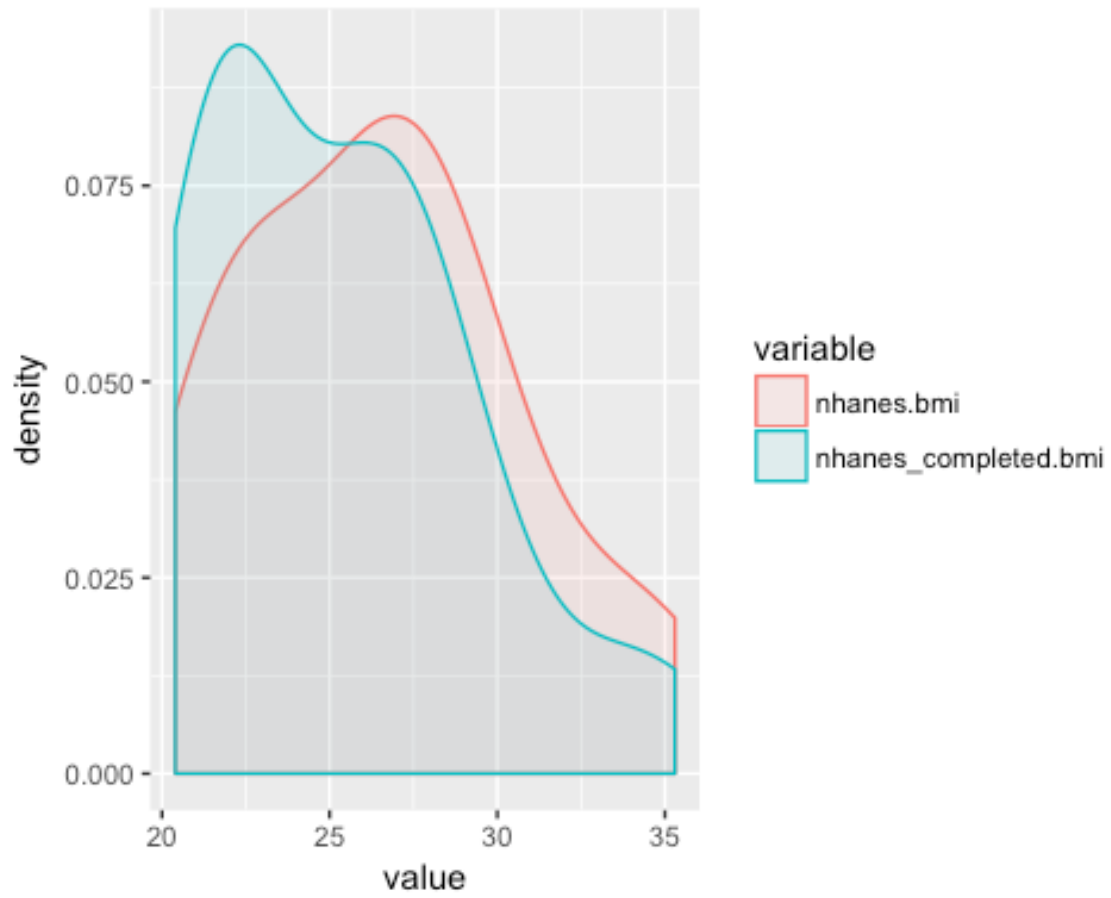
```



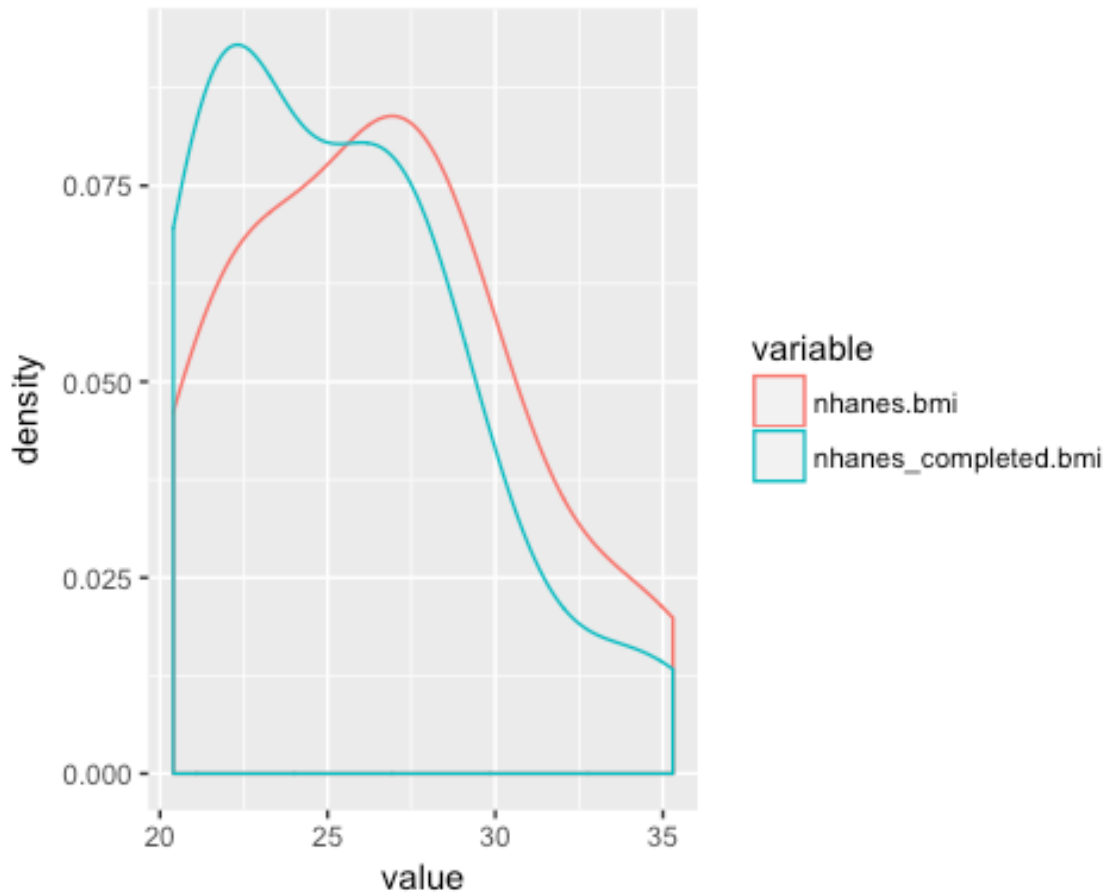

```
ggplot(comp_age, aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



```
# Dataframe con las 2 variables bmi de Los 2 datasets  
comp_bmi <- data.frame(nhanes$bmi, nhanes_completed$bmi)  
comp_bmi <- melt(comp_bmi)  
  
## No id variables; using all as measure variables  
  
# Plot de densidad con las 2 variables bmi de Los 2 datasets  
ggplot(comp_bmi, aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable, fill=variable),  
    alpha=0.1)
```



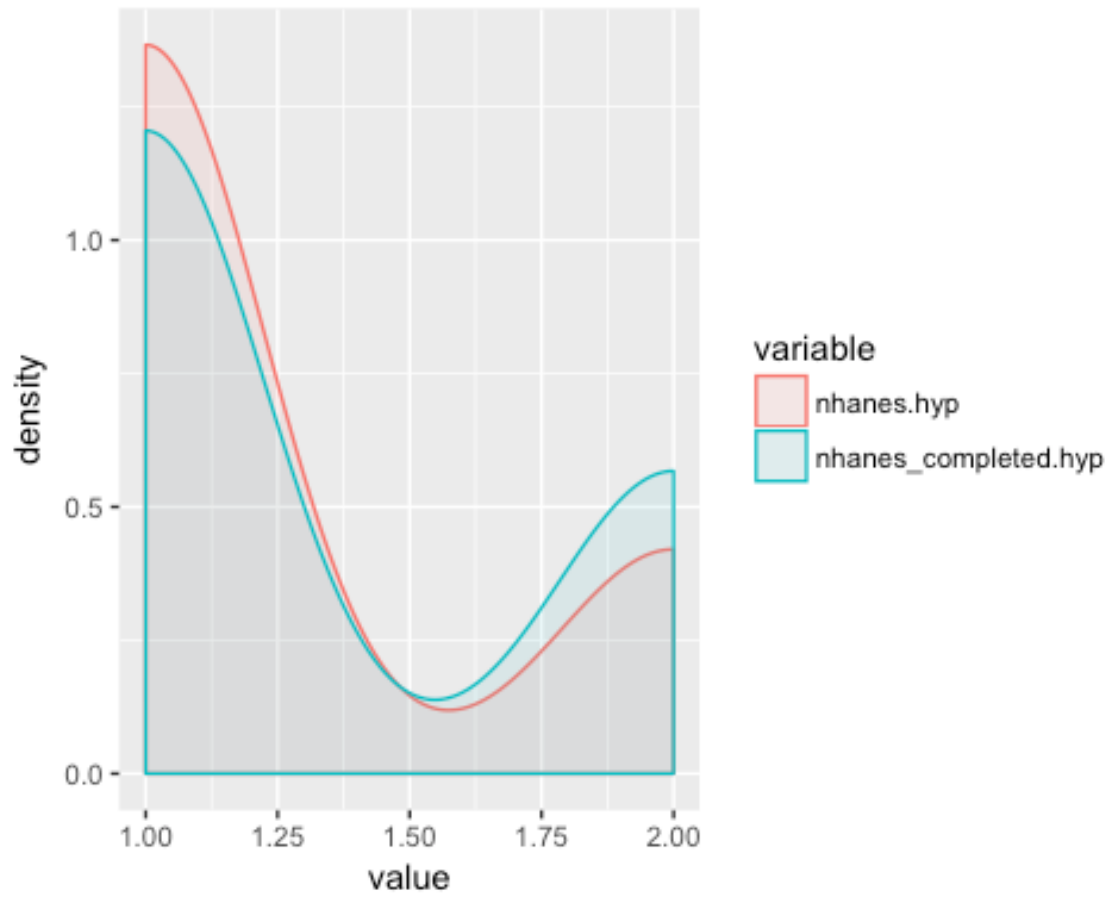
```
ggplot(comp_bmi, aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



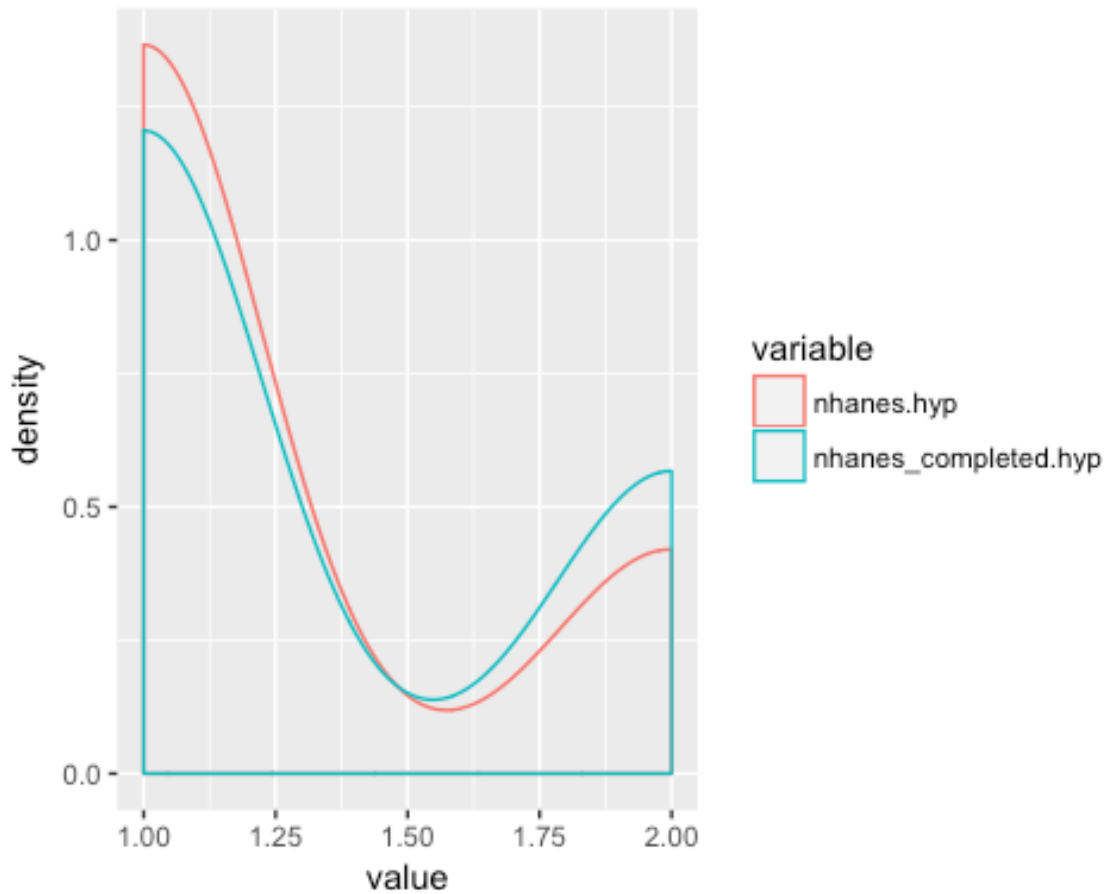
```
# Dataframe con las 2 variables hyp de los 2 datasets
comp_hyp <- data.frame(nhanes$hyp, nhanes_completed$hyp)
comp_hyp <- melt(comp_hyp)

## No id variables; using all as measure variables

# Plot de densidad con las 2 variables hyp de los 2 datasets (no se
# produjeron imputaciones porque hyp no tenía NAs)
ggplot(comp_hyp, aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
    alpha=0.1)
```



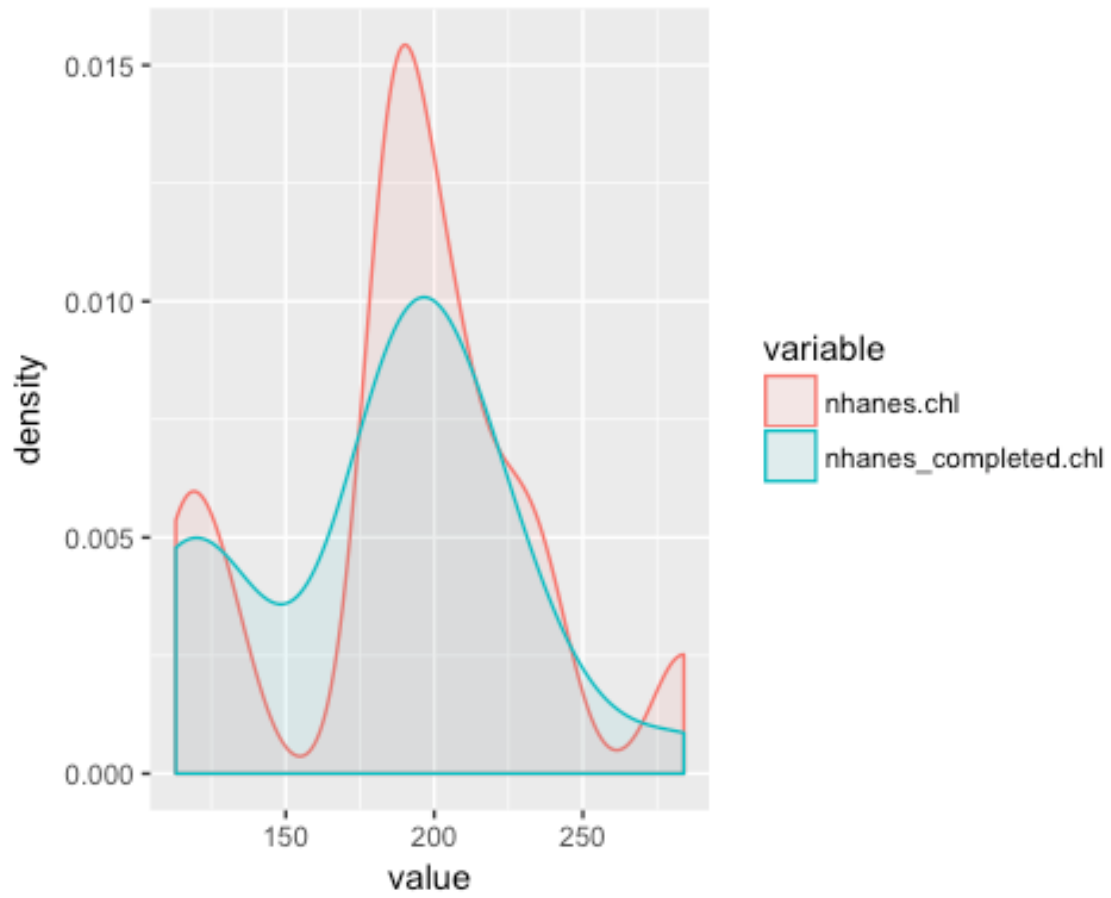
```
ggplot(comp_hyp , aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



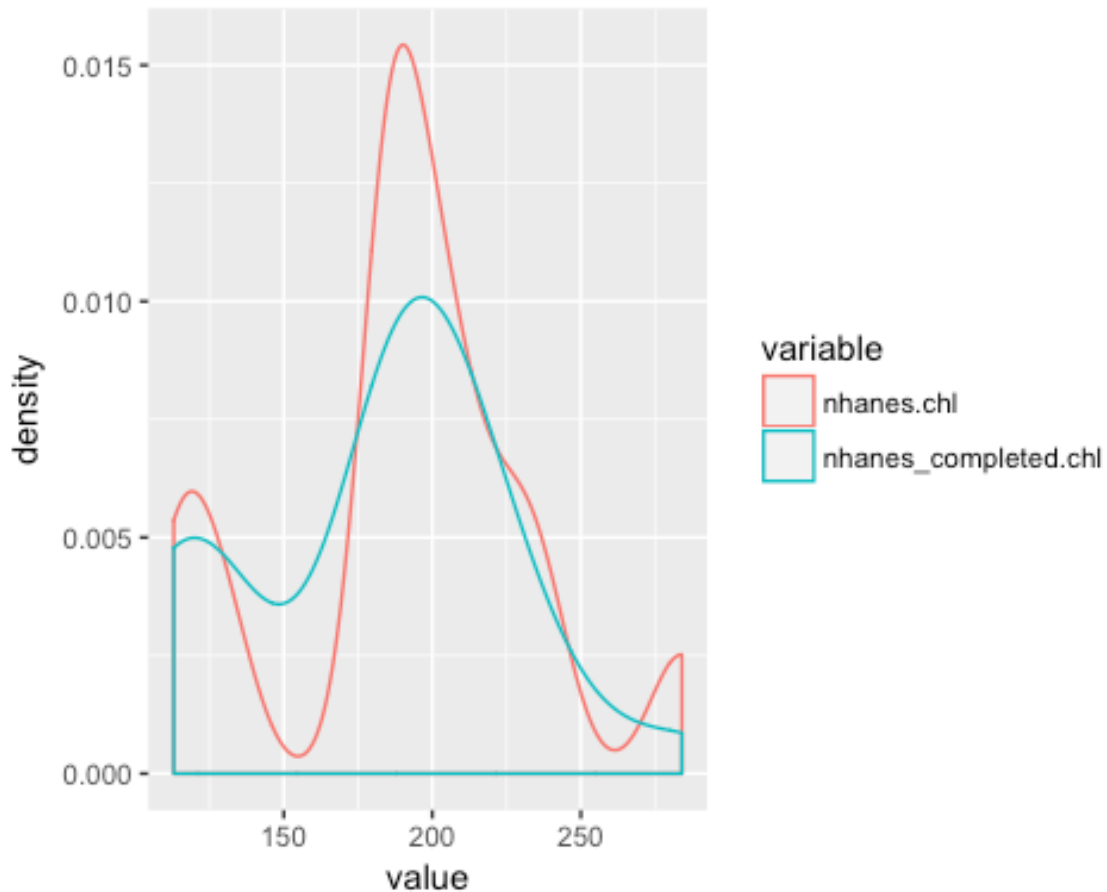
```
# Dataframe con las 2 variables chl de los 2 datasets
comp_chl <- data.frame(nhanes$chl, nhanes_completed$chl)
comp_chl <- melt(comp_chl)

## No id variables; using all as measure variables

# Plot de densidad con las 2 variables chl de los 2 datasets (no se
produjeron imputaciones porque Wind no tenía NAs)
ggplot(comp_chl , aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)
```



```
ggplot(comp_ch1 , aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



4.5 Dataset "BOYS"

```
# Dataset BOYS
```

```
data(boys)
```

```
# Estructura del dataset
```

```
str(boys)
```

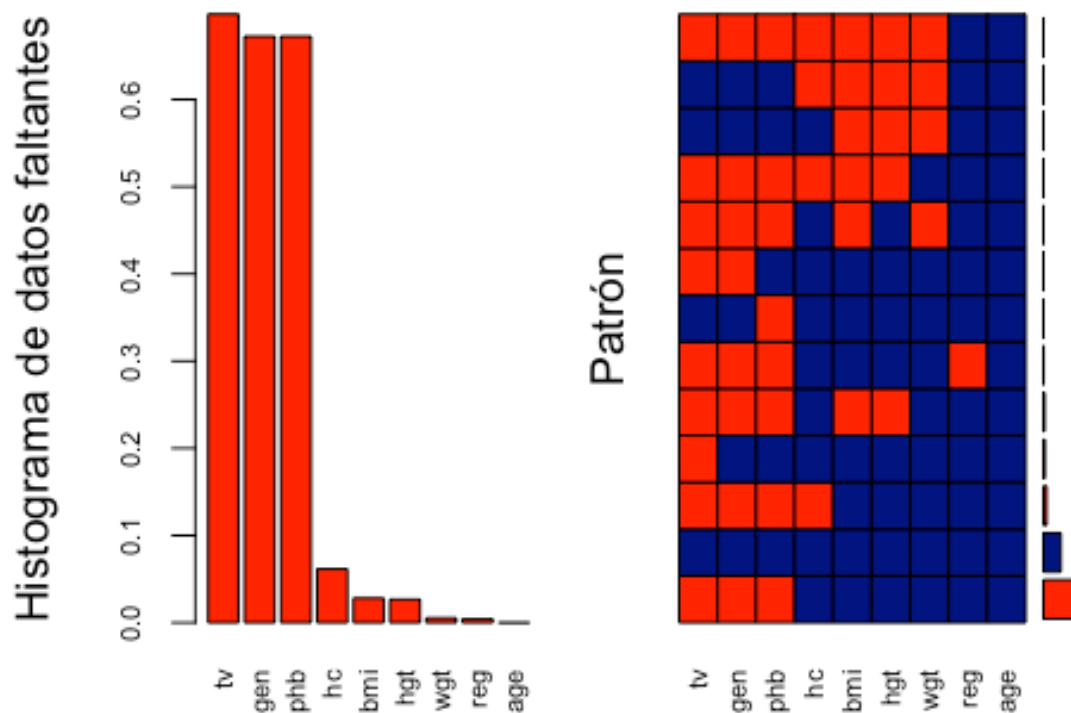
```
## 'data.frame':  748 obs. of  9 variables:
## $ age: num  0.035 0.038 0.057 0.06 0.062 0.068 0.068 0.071 0.071
##      0.073 ...
## $ hgt: num  50.1 53.5 50 54.5 57.5 55.5 52.5 53 55.1 54.5 ...
## $ wgt: num   3.65 3.37 3.14 4.27 5.03 ...
## $ bmi: num   14.5 11.8 12.6 14.4 15.2 ...
## $ hc : num   33.7 35 35.2 36.7 37.3 37 34.9 35.8 36.8 38 ...
## $ gen: Ord.factor w/ 5 levels "G1"<"G2"<"G3"<...: NA NA NA NA NA NA NA
##      NA NA NA ...
## $ phb: Ord.factor w/ 6 levels "P1"<"P2"<"P3"<...: NA NA NA NA NA NA NA
##      NA NA NA ...
```

```
## $ tv : int  NA NA NA NA NA NA NA NA NA NA NA ...
## $ reg: Factor w/ 5 levels "north","east",...: 4 4 4 4 4 4 4 4 3 3 2 ...

# Guardo el dataset BOYS para generar el fichero BOYS.rds
# (a efectos de reproducibilidad posterior)
saveRDS(boys, "BOYS.rds")

# Apertura del dataset BOYS
# (Descomentar la línea siguiente si fuera necesario abrirlo para
# reproducibilidad posterior)
# boys <- readRDS("BOYS.rds")

# Ilustración 25: Visualización de patrón de datos faltantes del dataset
# BOYS
aggr_plot <- aggr(boys, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(boys), cex.axis=.7, gap=3, ylab=c("Histograma
de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable      Count
##      tv 0.697860963
##      gen 0.672459893
```

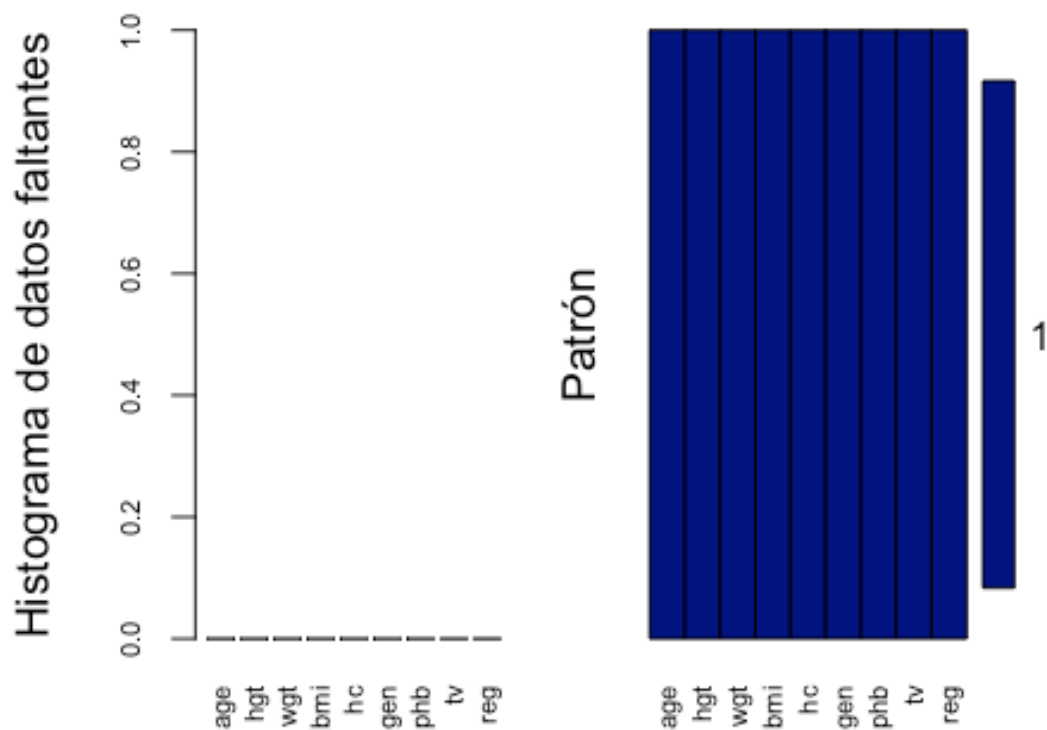


```
##      phb 0.672459893
##      hc 0.061497326
##      bmi 0.028074866
##      hgt 0.026737968
##      wgt 0.005347594
##      reg 0.004010695
##      age 0.000000000

# Imputación múltiple del dataframe BOYS
imp <- mice(boys, seed=5, print=FALSE)

# Imputación efectiva del dataset BOYS
boys_completed = complete(imp)

# Ilustración XX: Visualización de patrón de datos en dataset BOYS
después de la imputación efectiva
aggr_plot <- aggr(boys_completed, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(boys_completed), cex.axis=.7, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable Count
```

```
##      age      0
##      hgt      0
##      wgt      0
##      bmi      0
##      hc       0
##      gen      0
##      phb      0
##      tv       0
##      reg      0

# Verificación del modelo de imputación empleado en cada variable
imp$meth

##      age      hgt      wgt      bmi      hc      gen      phb
##      ""      "pmm"    "pmm"    "pmm"    "pmm"    "polr"    "polr"
##      tv      reg
##      "pmm"  "polyreg"

# Comparación de dataset original con dataset imputado
head(boys)

##      age hgt  wgt  bmi  hc  gen  phb tv  reg
## 3  0.035 50.1 3.650 14.54 33.7 <NA> <NA> NA south
## 4  0.038 53.5 3.370 11.77 35.0 <NA> <NA> NA south
## 18 0.057 50.0 3.140 12.56 35.2 <NA> <NA> NA south
## 23 0.060 54.5 4.270 14.37 36.7 <NA> <NA> NA south
## 28 0.062 57.5 5.030 15.21 37.3 <NA> <NA> NA south
## 36 0.068 55.5 4.655 15.11 37.0 <NA> <NA> NA south

head(boys_completed) # Imputado con PMM

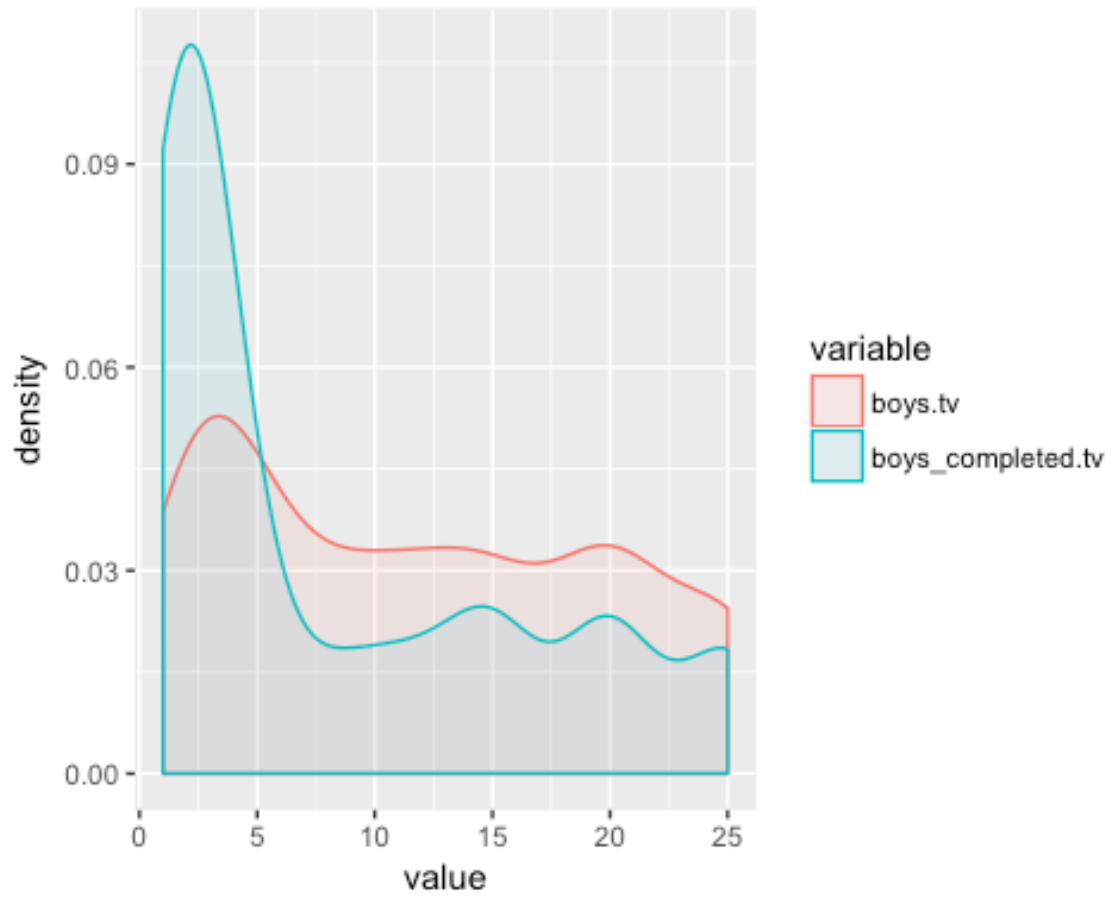
##      age hgt  wgt  bmi  hc  gen  phb tv  reg
## 3  0.035 50.1 3.650 14.54 33.7  G1  P2  1  south
## 4  0.038 53.5 3.370 11.77 35.0  G1  P3  2  south
## 18 0.057 50.0 3.140 12.56 35.2  G4  P5  2  south
## 23 0.060 54.5 4.270 14.37 36.7  G2  P1  2  south
## 28 0.062 57.5 5.030 15.21 37.3  G1  P1  2  south
## 36 0.068 55.5 4.655 15.11 37.0  G3  P3  2  south

# Dataframe con las 2 variables tv de los 2 datasets
comp_tv <- data.frame(boys$tv, boys_completed$tv)
comp_tv <- melt(comp_tv)

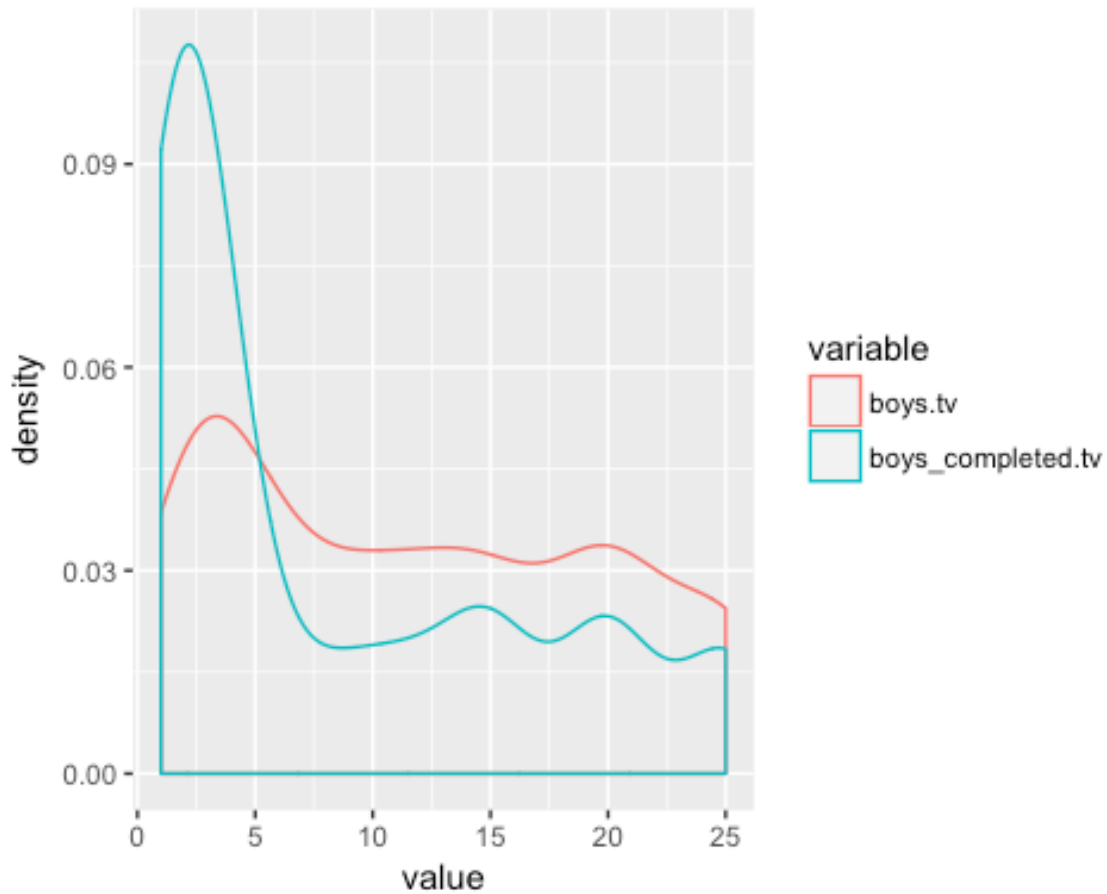
## No id variables; using all as measure variables

# Cargo librería GGPlot2
library(ggplot2)

# Plot de densidad con las 2 variables tv de los 2 datasets
ggplot(comp_tv, aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)
```



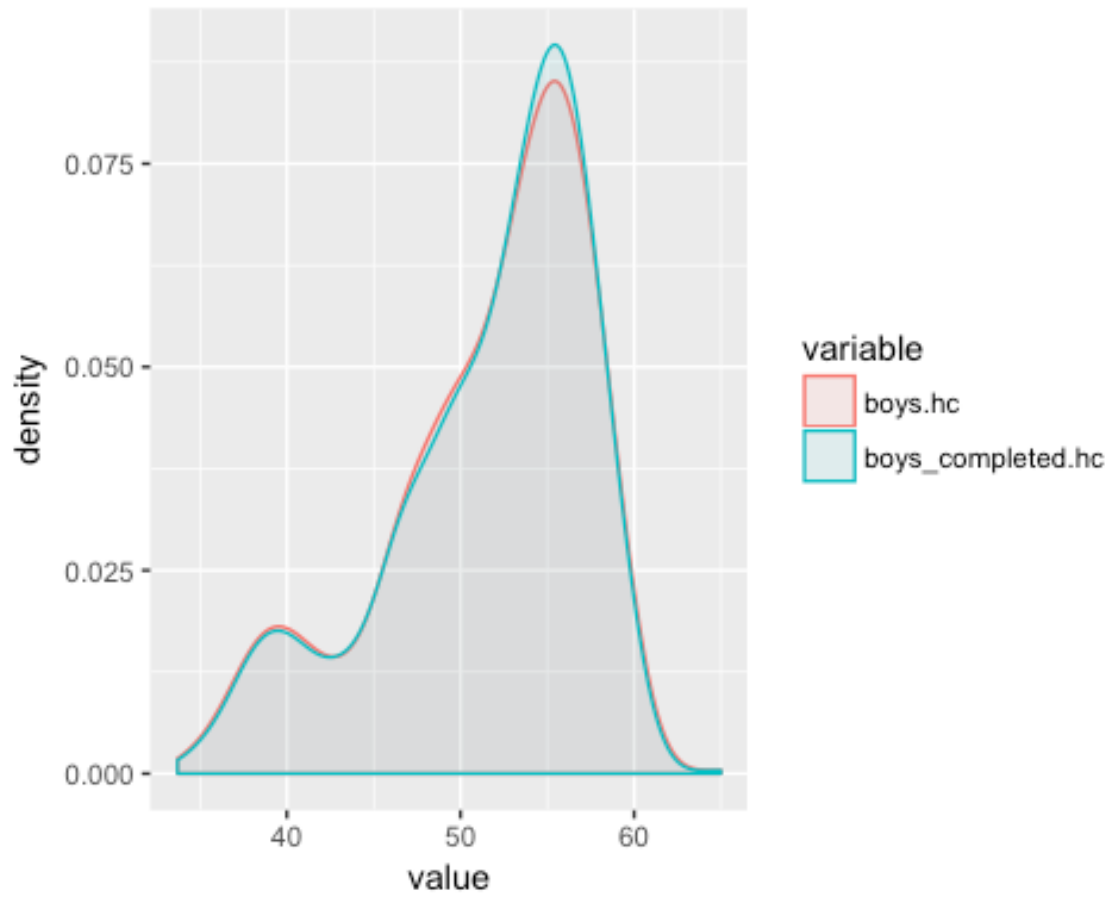
```
ggplot(comp_tv, aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



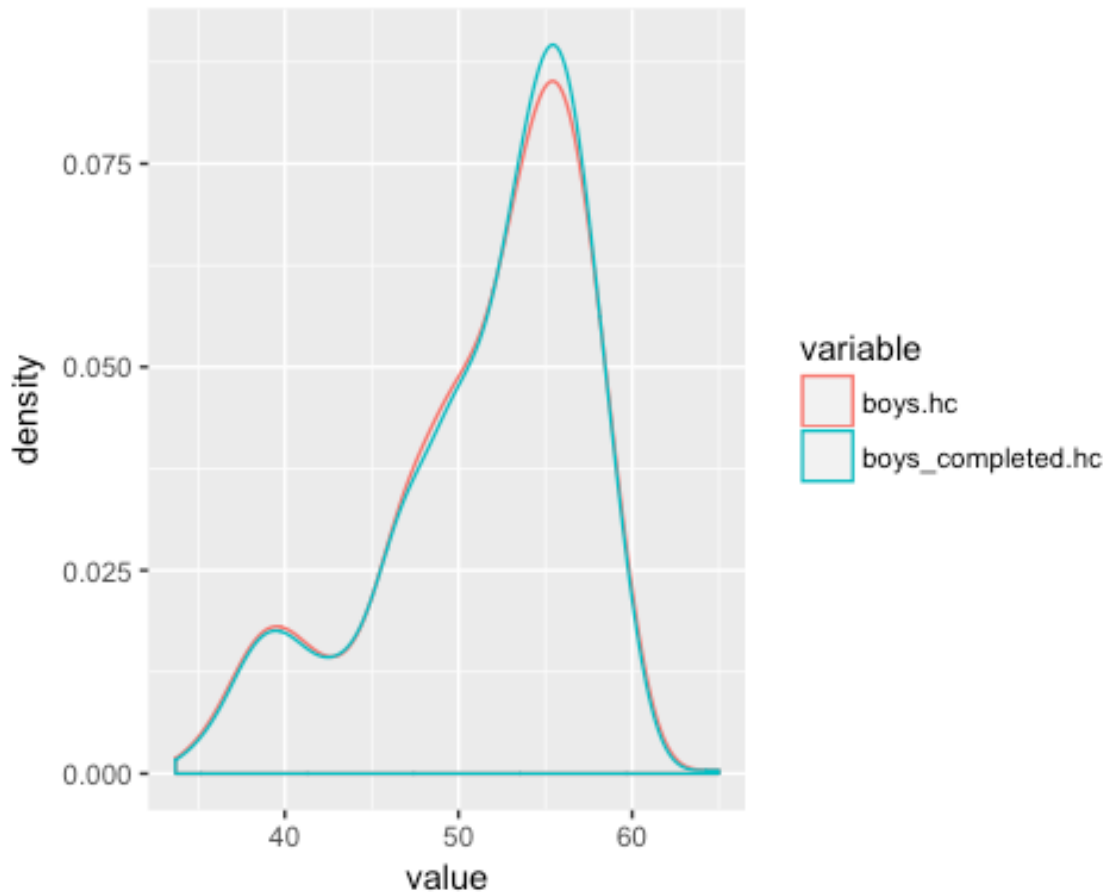
```
# Dataframe con Las 2 variables hc de Los 2 datasets
comp_hc <- data.frame(boys$hc, boys_completed$hc)
comp_hc <- melt(comp_hc)

## No id variables; using all as measure variables

# Plot de densidad con Las 2 variables hc de Los 2 datasets
ggplot(comp_hc , aes(x=value)) +
  geom_density(aes(group=variable, colour=variable, fill=variable),
alpha=0.1)
```



```
ggplot(comp_hc , aes(x=value)) +  
  geom_density(aes(group=variable, colour=variable))
```



4.6 Dataset "WALK"

```
# Dataset WALK (Walking)
data(walking)

# Estructura del dataset
str(walking)

## 'data.frame':    890 obs. of  5 variables:
## $ sex: Factor w/ 2 levels "Male","Female": 1 2 1 1 2 1 2 2 2 2 ...
## $ age: num  61 69 74 66 72 67 66 66 64 57 ...
## $ YA : Ord.factor w/ 4 levels "0"<"1"<"2"<"3": 2 2 1 1 3 1 1 2 1 1
## $ YB : Ord.factor w/ 4 levels "0"<"1"<"2"<"3": NA NA NA NA NA NA NA
## $ src: Factor w/ 3 levels "A","B","E": 1 1 1 1 1 1 1 1 1 1 ...

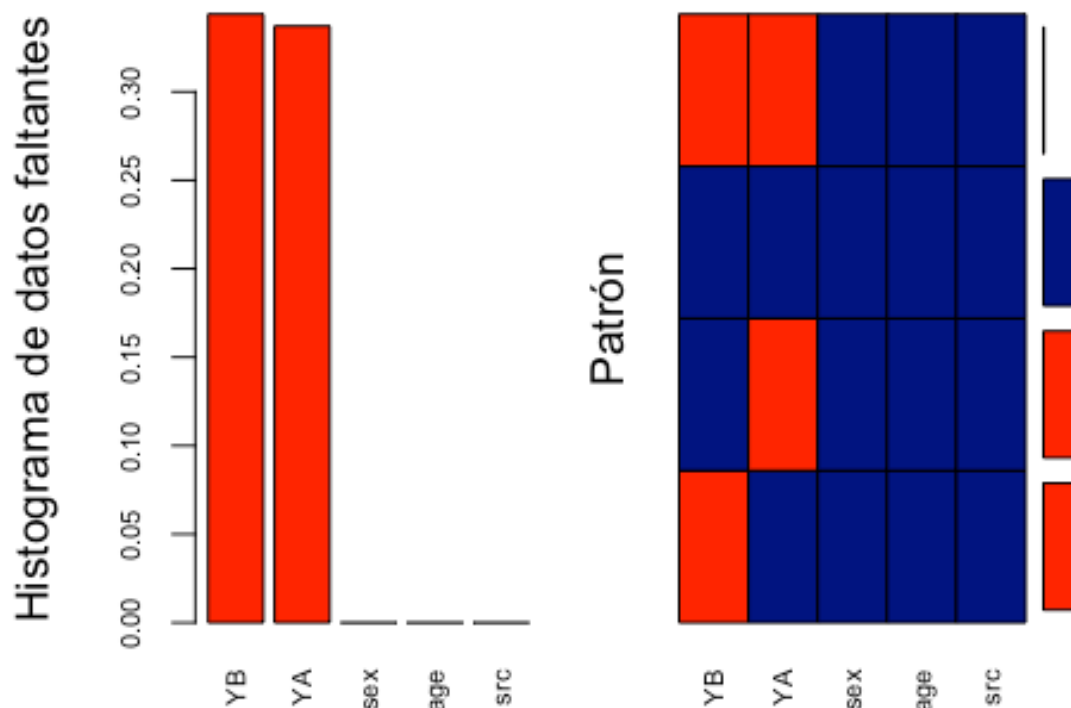
# Guardo el dataset WALK para generar el fichero WALK.rds
# (a efectos de reproducibilidad posterior)
saveRDS(walking, "WALK.rds")
```

```

# Apertura del dataset WALK
# (Descomentar la línea siguiente si fuera necesario abrirlo para
reproducibilidad posterior)
# walking <- readRDS("WALK.rds")

# Ilustración 26: Visualización de patrón de datos faltantes del dataset
WALK
aggr_plot <- aggr(walking, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(walking), cex.axis=.7, gap=3,
ylab=c("Histograma de datos faltantes", "Patrón"))

```



```

##
## Variables sorted by number of missings:
## Variable      Count
##      YB 0.3438202
##      YA 0.3370787
##      sex 0.0000000
##      age 0.0000000
##      src 0.0000000

```

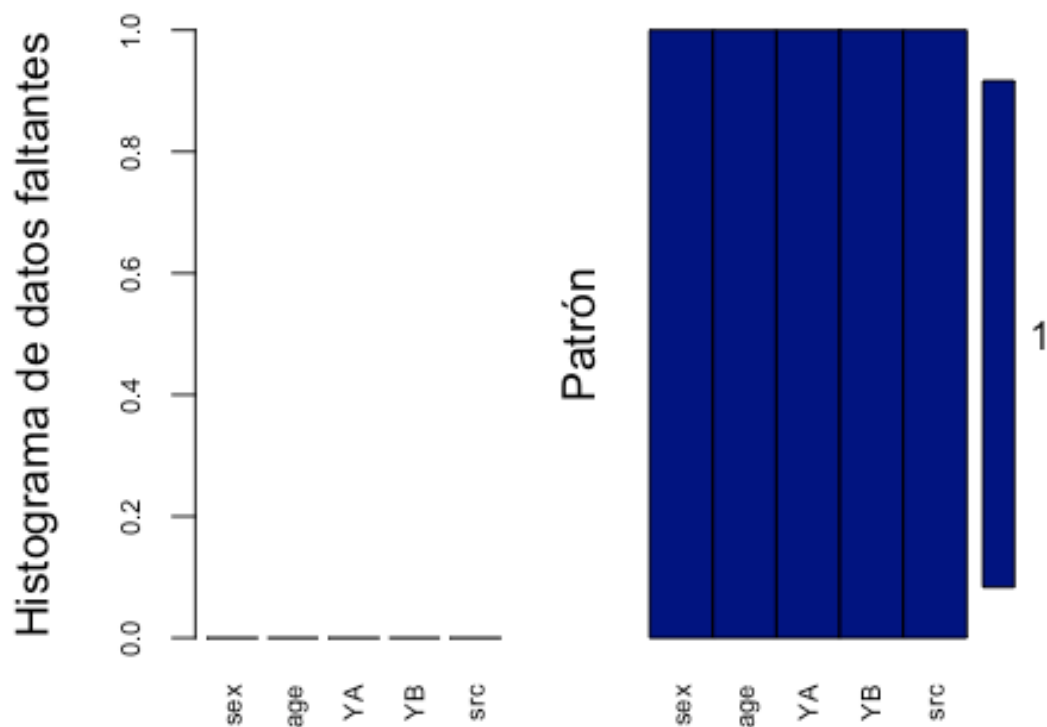
```

# Imputación múltiple del dataframe WALK
imp <- mice(walking, seed=5, print=FALSE)

```

```
# Imputación efectiva del dataset WALK
walking_completed = complete(imp)

# Ilustración XX: Visualización de patrón de datos en dataset WALK
después de la imputación efectiva
aggr_plot <- aggr(walking_completed, col=c('navyblue','red'),
numbers=TRUE, sortVars=TRUE, labels=names(walking_completed),
cex.axis=.7, gap=3, ylab=c("Histograma de datos faltantes", "Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable Count
##      sex      0
##      age      0
##      YA       0
##      YB       0
##      src      0
```

```
# Verificación del modelo de imputación empleado en cada variable
imp$meth
```



```
##      sex      age      YA      YB      src
##      ""      ""    "polr" "polr"      ""
```

4.7 Dataset "PATT"

```
# Dataset PATT
data(pattern4)

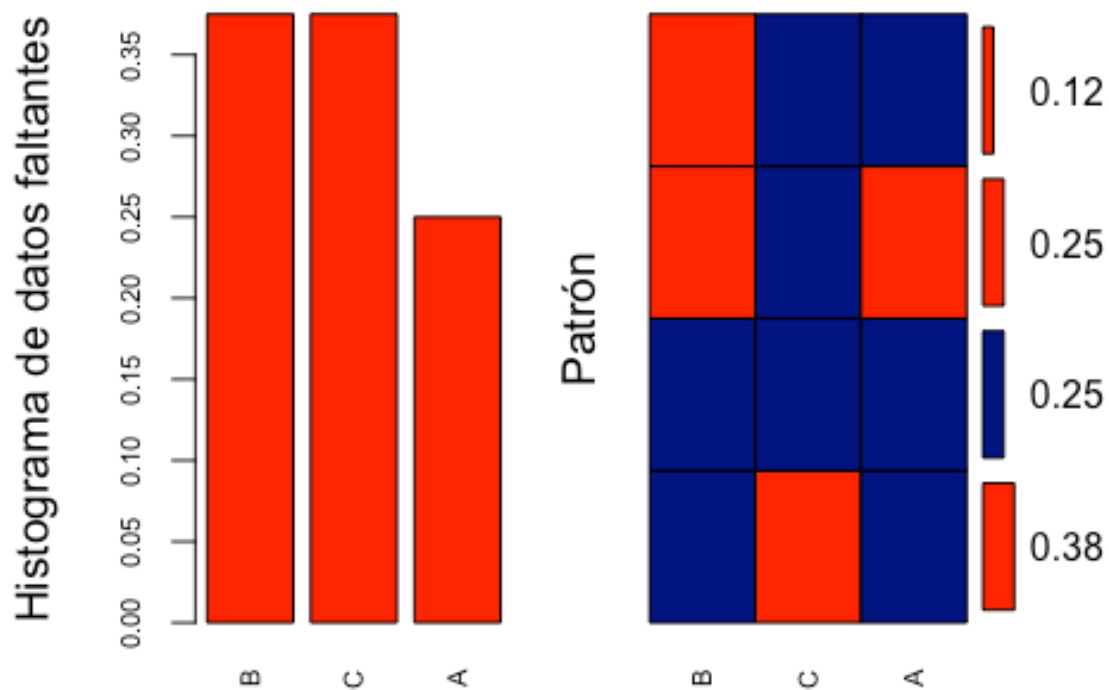
# Estructura del dataset
str(pattern4)

## 'data.frame':    8 obs. of  3 variables:
##  $ A: int  26 42 86 9 20 89 NA NA
##  $ B: int  88 66 54 92 83 NA NA NA
##  $ C: int  32 21 NA NA NA 41 35 33

# Guardo el dataset PATT para generar el fichero PATT.rds
# (a efectos de reproducibilidad posterior)
saveRDS(pattern4, "PATT.rds")

# Apertura del dataset PATT
# (Descomentar la línea siguiente si fuera necesario abrirlo para
# reproducibilidad posterior)
# pattern4 <- readRDS("PATT.rds")

# Ilustración 28: Visualización de patrón de datos faltantes del dataset
PATT
aggr_plot <- aggr(pattern4, col=c('navyblue','red'), numbers=TRUE,
sortVars=TRUE, labels=names(pattern4), cex.axis=.7, gap=3,
ylab=c("Histograma de datos faltantes","Patrón"))
```

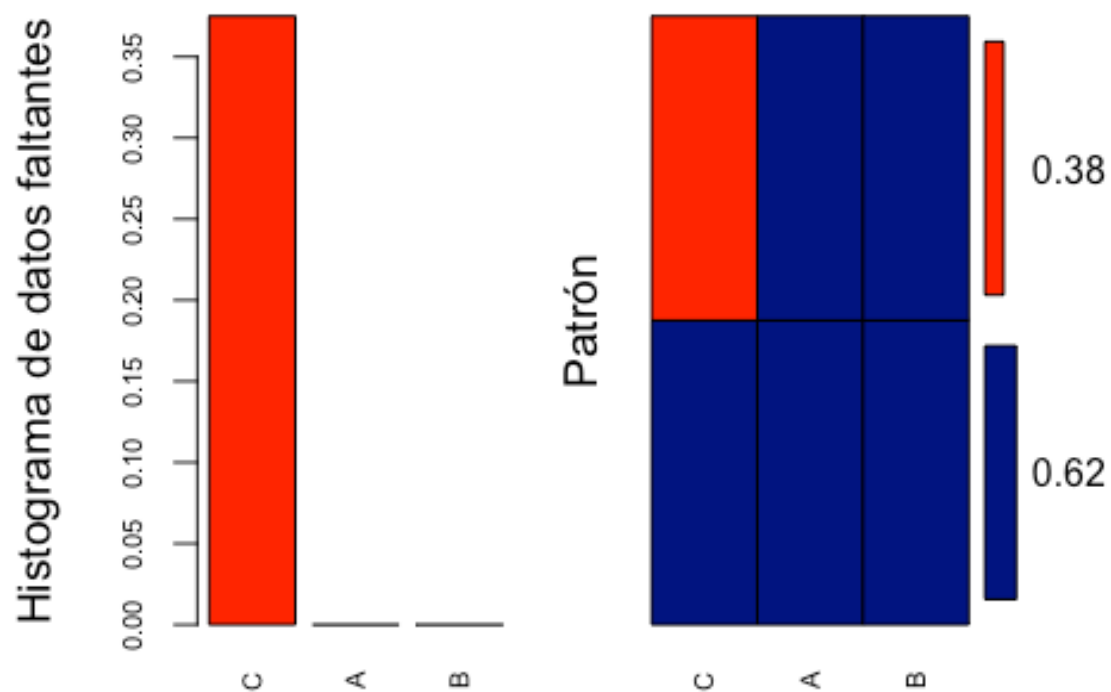


```
##
## Variables sorted by number of missings:
## Variable Count
##      B 0.375
##      C 0.375
##      A 0.250

# Imputación múltiple del dataframe PATT
imp <- mice(pattern4, seed=4, print=FALSE)

# Imputación efectiva del dataset PATT
pattern4_completed = complete(imp)

# Ilustración XX: Visualización de patrón de datos en dataset PATT
después de la imputación efectiva
aggr_plot <- aggr(pattern4_completed, col=c('navyblue','red'),
numbers=TRUE, sortVars=TRUE, labels=names(pattern4_completed),
cex.axis=.7, gap=3, ylab=c("Histograma de datos faltantes","Patrón"))
```



```
##
## Variables sorted by number of missings:
## Variable Count
##      C 0.375
##      A 0.000
##      B 0.000

# Verificación del modelo de imputación empleado en cada variable
imp$meth

##      A      B      C
## "pmm" "pmm" "pmm"
```