A1 - Preprocés de dades

Solució

Semestre 2022.1

Índex

Introducció

El conjunt de dades està a l'arxiu `gpa_row.csv`, conté la nota mitjana d'estudiants universitaris després del primer semestre de classes (GPA: grade point average, en anglès), així com informació sobre la nota d'accés, la cohort de graduació a l'institut i algunes característiques dels estudiants.

Aquest conjunt de dades surt d'una enquesta realitzada a una mostra representativa d'estudiants d'una universitat dels EUA (per raons de confidencialitat el conjunt de dades no inclou el nom de la universitat). Les variables incloses al conjunt de dades són:

- sat: nota d'accés (escala de 400 a 1600 punts)
- tothrs: hores totals cursades al semestre
- hsize: nombre total d'estudiants a la cohort de graduats del batxillerat (en centenars)
- hsrank: rànquing de l'estudiant, donat per la nota mitjana del batxillerat, en la cohort de graduats del batxillerat
- hsperc: rànquing relatiu de l'estudiant (hsrank/hsize)
- colgpa: nota mitjana de l'estudiant al final del primer semestre (escala de 0 a 4 punts)
- athlete: indicador de si l'estudiant practica algun esport a la universitat
- female: indicador de si l'estudiant és dona
- white: indicador de si l'estudiant és de raça blanca o no
- black: indicador de si l'estudiant és de raça negra o no

L'objectiu d'aquesta activitat és preparar el fitxer per a la seva posterior anàlisi. Per a això, s'examinarà el fitxer per detectar i corregir possibles errors, inconsistències i valors perduts. A més, es presentarà una breu estadística descriptiva amb gràfics.

Per altra part, es farà un informe executiu que resumirà el que s'ha fet. Constarà de dues parts:

1. Documentar tots els canvis realitzats a les dades originals.
2. Breu resum de les característiques més destacables de cada variable.

Criteris de verificació i de normalització de les variables:

A continuació es mostren els criteris amb els que s'han de netejar les dades del conjunt:

1. Verificar que les variables de tipus indicador han de tenir només el valor TRUE o FALSE (majúscules i sense espais en blanc) i s'han de codificar com a variables categòriques ("factor"). En cas que no es compleixi, cal corregir-ho.
2. En les dades de naturalesa numèriques, el símbol de separador decimal és el punt i no la coma. A més, si es presenta la unitat de la variable cal eliminar-la per convertir la variable a tipus numèric.
3. Comprovar si es compleix el rang de valors possibles a les variables on es té aquesta informació:
    - 'sat' : escala de 400 a 1600 punts
    - 'colgpa' : escala de 0 a 4 punts
4. Revisar si els valors de la variable 'hsperc' s'ha calculat correctament a partir de 'hsrank / hsize' amb tres decimals de precisió. En cas contrari, modificar-ho.

Nota important a tenir en compte per a lliurar l'activitat:

- És necessari lliurar l'arxiu Rmd i el fitxer de sortida (PDF o html). L'arxiu de sortida ha d'incloure: el codi i el resultat de l'execució del codi (pas a pas).
- S'ha de respectar la mateixa numeració dels apartats que l'enunciat.

- No es poden realitzar llistats complets del conjunt de dades en la solució. Això generaria un document amb centenars de pàgines i dificulta la revisió del text. Per a comprovar les funcionalitats del codi sobre les dades, es poden usar les funcions **head** i **tail** que només mostren unes línies del fitxer de dades.
- Es valora la precisió dels termes utilitzats (cal fer servir de manera precisa la terminologia de l'estadística).
- Es valora també la concisió en la resposta. No es tracta de fer explicacions molt llargues o documents molt extensos. Cal explicar el resultat i argumentar la resposta a partir dels resultats obtinguts de manera clara i concisa.

Per realitzar el preprocés del fitxer, seguiu els passos que s'indiquen a continuació.

---

1 Càrrega de l'arxiu

Carregueu el fitxer de dades i examineu el tipus de dades amb què R ha interpretat cada variable.

Indiqueu quines variables són de naturalesa numèrica, tot i que R ho hagi pogut interpretar de manera diferent. En el cas que el tipus de variable que ha atorgat R no coincideixi amb el tipus que li correspondria, haureu d'aplicar la transformació corresponent quan feu la normalització de la variable (apartat següent).

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

     return (ds)
}
```

```
ds<-read.csv("gpa_row.csv",stringsAsFactors=TRUE)

# Obtenim les dimensions del conjunt de dades, l'estructura i contingut.
dim(ds)
```

```
## [1] 4137   10
```

```
str(ds)
```

```
## 'data.frame':    4137 obs. of  10 variables:
##  $ sat    : int  920 1170 810 940 1180 980 880 980 1240 1230 ...
##  $ tothrs : Factor w/ 125 levels "100h","101h",..: 67 46 42 64 46 16 103 79 46 45 ...
##  $ hsize  : Factor w/ 649 levels "0,30000001","0,40000001",..: 9 649 122 553 223 277 324 277 379 9 .
##  $ hsrank : int  4 191 42 252 86 41 161 101 161 3 ...
##  $ hsperc : num  40 20.3 35.3 44.1 40.2 ...
##  $ colgpa : num  2.04 4 1.78 2.42 2.61 ...
##  $ athlete: Factor w/ 4 levels "false","FALSE",..: 4 2 4 2 2 2 2 2 2 2 ...
##  $ female : logi  TRUE FALSE FALSE FALSE FALSE TRUE ...
##  $ white  : Factor w/ 6 levels "  TRUE","false",..: 3 5 5 5 5 5 3 5 5 5 ...
##  $ black  : Factor w/ 6 levels "  FALSE","false",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
head(ds)
```

```
##    sat tothrs     hsize hsrank   hsperc colgpa athlete female white black
## 1  920    43h       0.1      4 40.00000   2.04    TRUE   TRUE FALSE FALSE
## 2 1170    18h 9.3999996    191 20.31915   4.00   FALSE  FALSE  TRUE FALSE
## 3  810    14h 1.1900001     42 35.29412   1.78    TRUE  FALSE  TRUE FALSE
## 4  940    40h      5.71    252 44.13310   2.42   FALSE  FALSE  TRUE FALSE
```

```
## 5 1180    18h 2.1400001     86 40.18692   2.61   FALSE  FALSE  TRUE FALSE
## 6  980   114h 2.6800001     41 15.29851   3.03   FALSE   TRUE  TRUE FALSE
```

```
summary(ds)
```

```
##       sat           tothrs           hsize          hsrank
##  Min.   : 470   17h    : 305   0.1      : 115   Min.   :  1.00
##  1st Qu.: 940   16h    : 279   2.3399999:  49   1st Qu.: 11.00
##  Median :1030   15h    : 226   2.8      :  49   Median : 30.00
##  Mean   :1030   14h    : 167   2.1099999:  41   Mean   : 52.83
##  3rd Qu.:1120   18h    : 153   2.03     :  37   3rd Qu.: 70.00
##  Max.   :1540   13h    : 146   2.3800001:  36   Max.   :634.00
##                 (Other):2861   (Other)  :3810
##      hsperc            colgpa       athlete       female           white
##  Min.   : 0.1667   Min.   :0.000   false:  11   Mode :logical    TRUE :   2
##  1st Qu.: 6.4328   1st Qu.:2.210   FALSE:3932   FALSE:2277      false  :   3
##  Median :14.5963   Median :2.660   true :   1   TRUE :1860      FALSE  : 305
##  Mean   :19.2406   Mean   :2.655   TRUE : 193                   true   :   9
##  3rd Qu.:27.7108   3rd Qu.:3.120                                TRUE   :3814
##  Max.   :92.0000   Max.   :4.000                                TRUE   :   4
##                    NA's   :41
##       black
##    FALSE :   3
##  false   :  10
##  FALSE   :3890
##  FALSE   :   5
##  TRUE    : 228
##  TRUE    :   1
##
```

```
id.factor   <- c(7:10)
id.num      <- c(1:6)
var.factor  <- colnames(ds)[id.factor]
var.num     <- colnames(ds)[id.num]

info <- data.frame(id=1, row="",
                   message= paste0("n. row = ", nrow(ds),"; ",
                                   "n. col= ", ncol(ds), "; ",
                                   "n. var num. = ",  length(id.num),"; ",
                                   "n. var qualit. = ", length(id.factor)))
```

Haurien de ser variables qualitatives (factor): **athlete, female, white, black**

Haurien de ser variables quantitatives (numèriques): **sat, tothrs, hsize, hsrank, hsperc, colgpa**

2 Normalització de les variables qualitatives

2.1 Athlete

Normalitzar la variable `Athlete` segons les indicacions proporcionades.

```
  # Revisió variable
  table( ds$athlete )
```

```
##
## false FALSE  true  TRUE
##    11  3932     1   193
```

```
  #Reporting de canvis
  idx <- which( ds$athlete =="false")
  idx
```

```
##  [1]   65  385  720  939 1293 1312 2330 2400 2933 3193 3543
```

```
  info <- report(info, row=paste(idx,collapse=", "), "athlete: false  -> FALSE")

  idx <- which( ds$athlete =="true")
  idx
```

```
## [1] 876
```

```
  info <- report(info, row=paste(idx,collapse=", "), "athlete: true  -> TRUE")

  #---------------------------------
  # Convertir a majúscules
  ds$athlete <- str_to_upper(ds$athlete)

  # Convertir a factor
  ds$athlete <- factor(ds$athlete)

  # checking
  table( ds$athlete )
```

```
##
## FALSE  TRUE
##  3943   194
```

2.2 Female

Normalitzar la variable `Female` segons les indicacions proporcionades.

```
  # Revisió variable
  table( ds$female )
```

```
##
## FALSE  TRUE
##  2277  1860
```

```
  # Convertir a factor
  ds$female <- factor(ds$female)


  # Tot correcte
```

2.3 Black

Normalitzar la variable `Black` segons les indicacions proporcionades.

```
  # Revisió variable
  table( ds$black )
```

```
##
##    FALSE    false    FALSE FALSE        TRUE  TRUE
##        3       10     3890        5      228        1
```

```
  #Reporting de canvis
  idx <- grep('[[:space:]]', ds$black)
  length(idx)
```

```
## [1] 9
```

```
  info <- report(info, row=paste(idx,collapse=", "), "black: Eliminat espais en blanc")

  idx <- which( ds$black =="false")
  idx
```

```
##  [1] 1255 1785 2227 2424 2450 2913 3076 3102 3803 3868
```

```
  info <- report(info, row=paste(idx,collapse=", "), "black: false  -> FALSE")

  #---------------------------------
  # Eliminar espais en blanc
  ds$black <- str_trim(ds$black)
  table( ds$black )
```

```
##
## false FALSE  TRUE
##    10  3898   229
```

```
  # Convertir a majúscules
  ds$black <- str_to_upper(ds$black)

  # Convertir a factor
  ds$black <- factor(ds$black)

  # checking
  table( ds$black )
```

```
##
## FALSE  TRUE
##  3908   229
```

2.4 White

Normalitzar la variable `white` segons les indicacions proporcionades.

```
  # Revisió variable
  table( ds$white )
```

```
##
##    TRUE   false   FALSE    true    TRUE TRUE
##       2       3     305       9    3814       4
```

```
  #Reporting de canvis
  idx <-grep('[[:space:]]', ds$white)
  length(idx)
```

```
## [1] 6
```

```
  info <- report(info, row=paste(idx,collapse=", "), "white: Eliminat espais en blanc")

  idx <- which( ds$white =="false")
  idx
```

```
## [1] 1929 2922 3536
```

```
  info <- report(info, row=paste(idx,collapse=", "), "white: false  -> FALSE")

  idx <- which( ds$white =="true")
  idx
```

```
## [1]  461  922 1007 1947 2810 2969 3129 4029 4030
```

```
  info <- report(info, row=paste(idx,collapse=", "), "white: true  -> TRUE")

  #----------------------------------
  # Eliminar espais en blanc
  ds$white <- str_trim(ds$white)
  table( ds$white )
```

```
##
## false FALSE  true  TRUE
##     3   305     9  3820
```

```
  # Convertir a majúscules
  ds$white <- str_to_upper(ds$white)

  # Convertir a factor
  ds$white <- factor(ds$white)

  # checking
  table( ds$white )
```

```
##
## FALSE  TRUE
##   308  3829
```

3 Normalització de les variables quantitatives

Inspeccionar els valors de les dades quantitatives i realitzar les normalitzacions oportunes seguint els criteris especificats anteriorment. Aquestes normalitzacions tenen com a objectiu uniformitzar els formats. Si hi ha valors perduts o valors atípics, es tractaran més endavant.

Al realitzar aquestes normalitzacions, s'ha de demostrar que la normalització sobre cada variable ha donat el resultat esperat. Per tant, es recomana mostrar un fragment de l'arxiu de dades resultant o un sumari. Per evitar presentar tot el conjunt de dades, es pot mostrar una part d'ell mateix, amb les funcions **head** i/o **tail**.

Seguiu l'ordre dels apartats.

3.1 Nota d'accés

Reviseu el format de la variable `sat` i feu les revisions o transformacions oportunes segons els criteris especificats anteriorment.

```
head(ds$sat,8)
```

```
## [1]  920 1170  810  940 1180  980  880  980
```

```
# Comprovació

idx <- which(ds$sat <  400 | ds$sat >  1600)
#
```

```
# Tots els valors són correctes


summary(ds$sat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     470     940    1030    1030    1120    1540
```

3.2 Hores totals cursades al semestre

Reviseu el format de la variable `tothrs` i feu les transformacions oportunes segons els criteris especificats anteriorment. Si hi ha valors atípics, es tractaran més endavant.

```
head(ds$tothrs,8)
```

```
## [1] 43h  18h  14h  40h  18h  114h 78h  55h
## 125 Levels: 100h 101h 102h 103h 104h 105h 106h 107h 108h 109h 10h 110h ... 9h
```

```
ds$tothrs <- as.numeric( trimws( sub('h', "", ds$tothrs ) ) )
head(ds$tothrs,8)
```

```
## [1]  43  18  14  40  18 114  78  55
```

```
# Comprobación
class(ds$tothrs)
```

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

```
summary(ds$tothrs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   17.00   47.00   52.83   80.00  137.00
```

```
#Reporting de canvis
 info <- report(info, row="*", "tothrs: Eliminat el text h i s'ha convertit a variable numèrica")
```

3.3 Nota mitjana de l'estudiant al final del primer semestre

Reviseu el format de la variable `colgpa` i feu les revisions o transformacions oportunes segons els criteris especificats anteriorment.

```
head(ds$colgpa,8)
```

```
## [1] 2.04 4.00 1.78 2.42 2.61 3.03 1.84 3.05
```

```
# Comprovació

idx <- which(ds$colgpa <  0 | ds$colgpa >  4)
#
# Tots els valors són correctes


summary(ds$colgpa)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.210   2.660   2.655   3.120   4.000      41
```

3.4 Nombre total d'estudiants a la cohort de graduats del batxillerat

Reviseu el format de la variable **hsize** i feu les revisions o transformacions oportunes segons els criteris especificats anteriorment.

```
class(ds$hsize)
```

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

```
head( ds$hsize, 30)
```

```
##  [1] 0.1        9.3999996  1.1900001  5.71       2.1400001  2.6800001
##  [7] 3.1099999  2.6800001  3.6700001  0.1        3.3399999  3.5899999
## [13] 3.1800001  1.92       3.6900001  2.6600001  1.45       1.76
## [19] 3.8599999  3.8299999  1.0700001  2.1700001  2.3399999  4.6300001
## [25] 5.9499998  0.91000003 4.3600001  8.1199999  0.60000002 3.76
## 649 Levels: 0,30000001 0,40000001 0,73000002 0.029999999 ... 9.3999996
```

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

#Reporting de canvis
idx <- grep("\\,",ds$hsize)
#ds$hsize[idx]
info <- report(info, row=paste(idx,collapse=", "), "corregim la coma pel punt decimal")

#---------------------------------
#corregim la coma pel punt decimal
ds$hsize <- gsub("\\,","\\.", ds$hsize)
ds$hsize <- as.numeric( ds$hsize )

#ckecking
head( ds$hsize, 30)
```

```
##  [1] 0.10 9.40 1.19 5.71 2.14 2.68 3.11 2.68 3.67 0.10 3.34 3.59 3.18 1.92 3.69
## [16] 2.66 1.45 1.76 3.86 3.83 1.07 2.17 2.34 4.63 5.95 0.91 4.36 8.12 0.60 3.76
```

3.5 Rànquing relatiu de l'estudiant

Reviseu si la variable **hsperc** s'ha obtingut correctament segons els criteris especificats anteriorment.

```
idx <- which(round((ds$hsrank/ ds$hsize),3)
               !=  round(ds$hsperc,3))
idx
```

```
##  [1]  188  201  313  657  876 2489 3438 3441 3445 3537 3753 4091
```

```
#Reporting de canvis
info <- report(info, row=paste(idx,collapse=", "), "hsperc: recalculem el valors amb hsrank/hsize")


#---------------------------------
ds$hsperc[idx] <- round(ds$hsrank[idx]/ ds$hsize[idx],3)

# Comprovació
idx <- which(round((ds$hsrank/ ds$hsize),3)
               !=  round(ds$hsperc,3))
idx
```

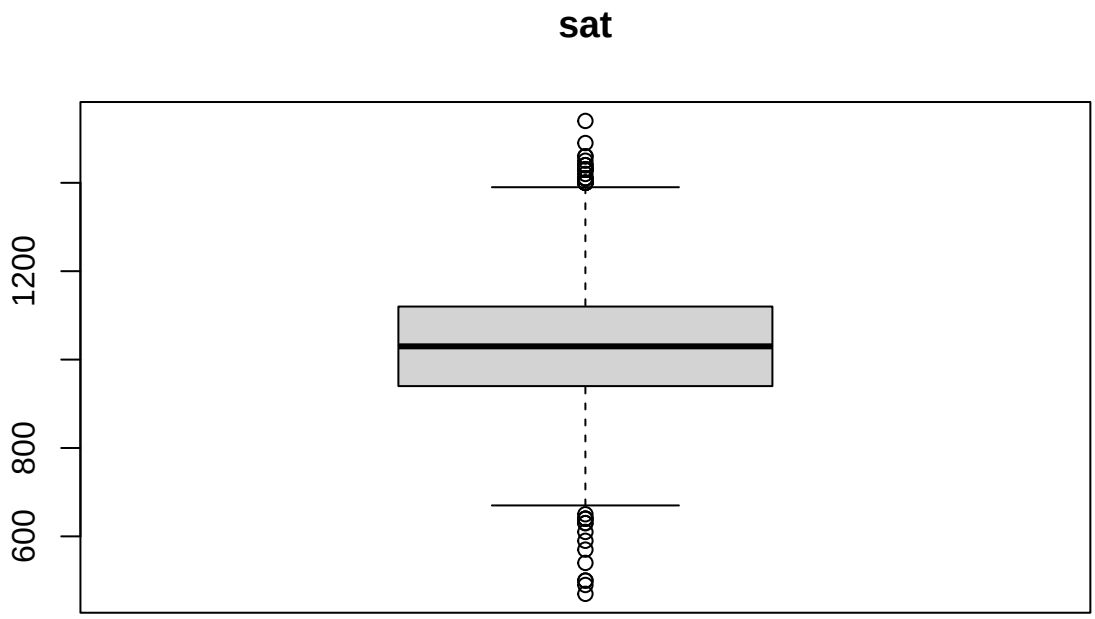
```
## integer(0)
```

```
# idx està buit
```

4 Valors atípics

Reviseu si hi ha valors atípics en les variables `sat` i `hsize`. Si es tracta d'un valor anòmal, és a dir anormalment alt o baix, substituir el seu valor per NA, que posteriorment s'ha d'imputar.

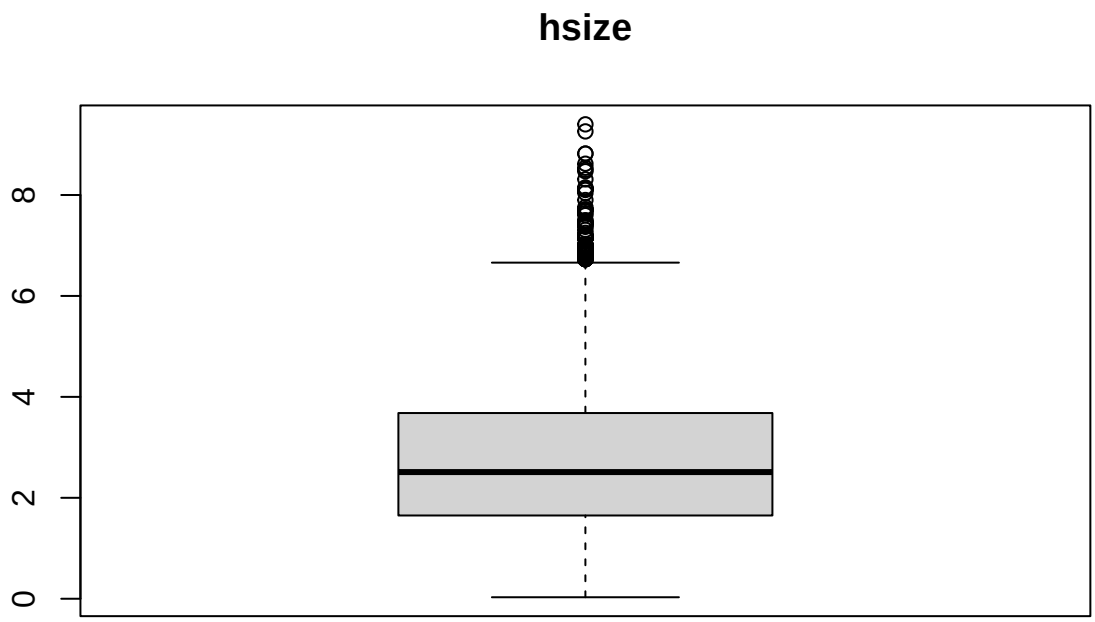
```
  #sat
  boxplot(ds$sat, main="sat")
```




```
  x<-boxplot.stats(ds$sat)$out
  idx <- which( ds$sat %in% x)
  sort(ds$sat[idx])
```

```
##  [1]  470  490  500  500  540  570  590  610  630  630  640  640  640  650 1400
## [16] 1400 1400 1400 1400 1400 1410 1410 1410 1410 1420 1430 1430 1430 1430 1430
## [31] 1430 1440 1440 1440 1450 1460 1460 1490 1540
```

```
  #Els valors són correctes. No es modifiquen.



  #hsize
  boxplot(ds$hsize, main="hsize")
```




```
  x<-boxplot.stats(ds$hsize)$out
  idx <- which( ds$hsize %in% x)
  sort(ds$hsize[idx])
```

```
##   [1] 6.73 6.73 6.73 6.73 6.73 6.73 6.73 6.73 6.73 6.75 6.75 6.75 6.75 6.77 6.80
##  [16] 6.82 6.82 6.82 6.82 6.82 6.82 6.86 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.87
##  [31] 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.87 6.90 6.90
##  [46] 6.93 6.93 6.95 6.95 6.97 6.97 6.97 6.97 6.97 6.98 6.98 6.98 6.98 6.98 6.98
##  [61] 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98
##  [76] 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 6.98 7.00 7.00 7.00 7.00 7.00
##  [91] 7.00 7.00 7.00 7.01 7.01 7.03 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15
## [106] 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15 7.15
## [121] 7.15 7.15 7.18 7.18 7.18 7.18 7.18 7.18 7.20 7.21 7.25 7.31 7.37 7.40 7.40
## [136] 7.42 7.45 7.45 7.45 7.46 7.49 7.50 7.60 7.64 7.68 7.71 7.71 7.71 7.71 7.76
## [151] 7.90 8.04 8.10 8.12 8.12 8.12 8.12 8.12 8.12 8.15 8.31 8.47 8.50 8.54 8.62
## [166] 8.82 8.82 9.26 9.40
```

```
  #Els valors són correctes. No es modifiquen.
```

5 Imputació de valors

Busqueu si hi ha valors perduts en les variables quantitatives. En el cas de detectar algun valor perdut cal realitzar una imputació de valors en aquestes variables. Apliqueu imputació per veïns més propers, utilitzant la distància de Gower, considerant en el còmput dels veïns més propers la resta de variables quantitatives. A més, considereu que la imputació s'ha de fer amb registres del mateix gènere. Per exemple, si un registre a imputar és dona, s'ha de realitzar la imputació usant només les variables quantitatives dels registres de dones.

Per realitzar aquesta imputació, podeu fer servir la funció "kNN" de la llibreria VIM amb un nombre de veïns igual a 11.

Mostreu que la imputació s'ha realitzat correctament, mostrant el resultat de les dades afectades per la imputació.

```
# total registres
nrow(ds)
```

```
## [1] 4137
```

```
# Número de valors NA a cada variable

rx <- colSums(is.na(ds))
rx
```

```
##     sat  tothrs   hsize  hsrank  hsperc  colgpa athlete  female   white   black
##       0       0       0       0       0      41       0       0       0       0
```

```
# Total sense valors NAs
idx <- complete.cases(ds)
# Registres no complets
which(!idx)
```

```
##  [1]   40  100  318  343  490  500  629  846 1053 1172 1226 1238 1319 1450 1605
## [16] 1866 1888 1937 1975 2035 2108 2184 2530 2536 2691 2721 2728 2879 3149 3196
## [31] 3495 3496 3523 3546 3651 3660 3758 3798 3943 3998 4015
```

```
#Reporting de canvis
info <- report(info,
               row=paste(which(!idx),collapse=", "),
               paste(sum(!idx),
                     "registres amb NA a la variable",
                     paste(names(which(rx>0)),collapse=", ")))


table(idx)
```

```
## idx
## FALSE  TRUE
##    41  4096
```

```
#Identifiquem per separat els NAs de gènere femení i els de gènere masculí
fem.idx <- which( is.na(ds$colgpa) & (ds$female=="TRUE") ); fem.idx
```

```
##  [1]  100  318  629 1172 1238 1319 1605 1866 1937 1975 2108 2530 2536 2721 2728
## [16] 3651 4015
```

```
mas.idx <- which( is.na(ds$colgpa) & ds$female=="FALSE"); mas.idx
```

```
##  [1]   40  343  490  500  846 1053 1226 1450 1888 2035 2184 2691 2879 3149 3196
## [16] 3495 3496 3523 3546 3660 3758 3798 3943 3998
```

```
#Imputem en els registres female=="TRUE"
new.ds.fem<- kNN( ds[ ds$female=="TRUE", var.num], variable="colgpa", k=11)

new.ds.fem[new.ds.fem$colgpa==TRUE,]
```

```
## [1] sat        tothrs     hsize      hsrank     hsperc     colgpa     colgpa_imp
## <0 rows> (or 0-length row.names)
```

Taula 1: imputació valors colgpa dones

	sat	tothrs	hsize	hsrank	hsperc	colgpa
100	1120	49	0.10	1	10.000000	3.31
318	1050	12	3.70	30	8.108109	2.60
629	860	80	6.55	100	15.267180	2.82
1172	990	82	3.62	20	5.524862	2.81
1238	1060	120	4.39	32	7.289294	3.15
1319	1100	82	5.11	61	11.937380	2.55
1605	1030	77	5.70	108	18.947371	2.59
1866	810	84	2.14	10	4.672897	2.78
1937	830	46	4.50	97	21.555559	2.37
1975	1000	72	0.81	34	41.975311	2.19
2108	970	47	2.68	37	13.805970	2.72
2530	970	78	3.38	7	2.071006	2.74
2536	940	15	3.10	19	6.129032	2.78
2721	920	101	1.20	12	10.000000	2.67
2728	490	127	1.44	55	38.194439	2.60
3651	910	68	0.53	42	79.245293	2.41
4015	890	80	3.77	194	51.458889	2.46

```
ds[fem.idx,]$colgpa <- new.ds.fem[new.ds.fem$colgpa_imp==TRUE,]$colgpa

kable(ds[fem.idx, var.num],
      caption="imputació valors colgpa dones")
```

```
#Imputem en els registres female=="FALSE"
new.ds.mas <- kNN( ds[ ds$female=="FALSE", var.num], variable="colgpa", k=11)
ds[mas.idx,]$colgpa <- new.ds.mas[new.ds.mas$colgpa_imp==TRUE,]$colgpa

kable(ds[mas.idx, var.num],
      caption="imputació valors colgpa homes")
```

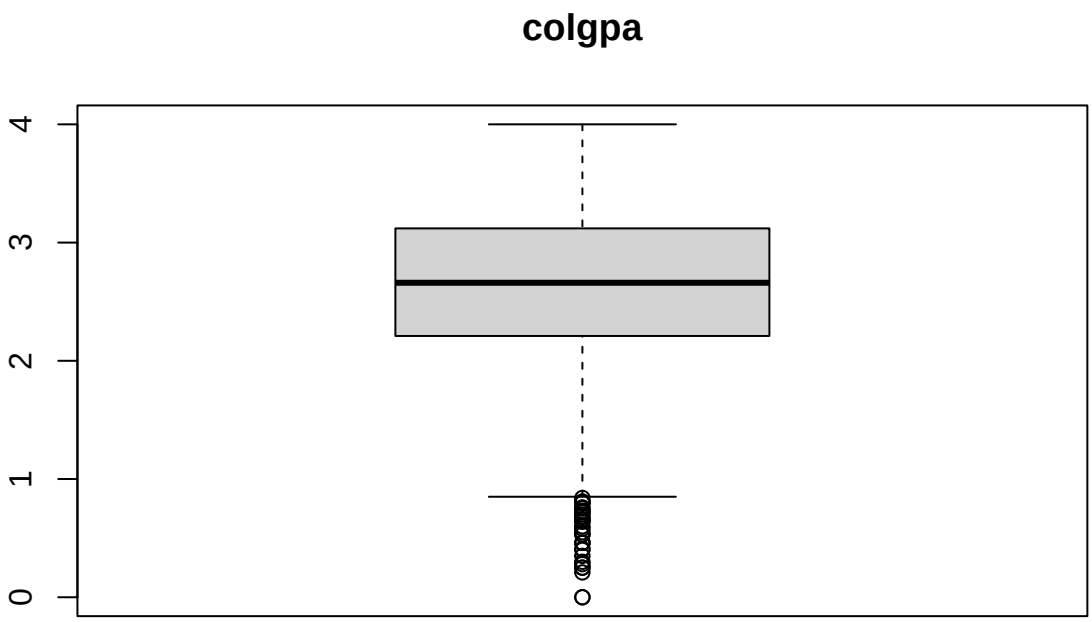
```
sum( complete.cases(ds$colgpa) )
```

```
## [1] 4137
```

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

Taula 2: imputació valors colgpa homes

	sat	tothrs	hsize	hsrank	hsperc	colgpa
40	940	19	1.85	41	22.162161	2.35
343	1100	43	7.45	218	29.261749	2.65
490	1090	95	4.72	67	14.194910	2.48
500	750	16	2.25	48	21.333330	2.18
846	970	39	2.21	77	34.841629	2.15
1053	900	17	1.54	44	28.571430	2.26
1226	1000	16	4.89	98	20.040899	2.72
1450	840	44	2.12	75	35.377361	2.26
1888	1040	16	0.83	23	27.710840	2.43
2035	990	78	0.72	3	4.166667	2.81
2184	1020	78	1.78	35	19.662920	2.60
2691	1260	50	7.00	47	6.714286	3.46
2879	1040	131	2.86	28	9.790210	2.69
3149	870	52	1.60	16	10.000000	2.50
3196	1070	14	2.34	36	15.384610	2.50
3495	910	13	4.89	145	29.652349	2.68
3496	910	40	4.77	125	26.205450	2.35
3523	1250	17	5.51	29	5.263158	2.76
3546	900	91	6.05	65	10.743800	2.42
3660	900	16	0.44	31	70.454536	2.26
3758	1160	120	1.72	11	6.395349	3.23
3798	930	44	2.92	6	2.054795	2.74
3943	1120	95	2.60	26	10.000000	3.41
3998	990	14	9.26	385	41.576679	1.68



6 Creació d'una nova variable

La variable `colgpa` conté la nota numèrica de l'alumnat. Crear una variable categòrica anomenada `gpaletter`, que indiqui la nota en lletra de cada estudiant de la següent forma: A, de 3.50 a 4.00; B, de 2.50 a 3.49; C, de 1.50 a 2.49; D, de 0 a 1.49.

```
gpanum <- ds$colgpa
gpa_level<-c("D","C","B", "A")
classif <- ifelse( gpanum<=1.49, gpa_level[1],
                   ifelse(gpanum<=2.49, gpa_level[2],
                   ifelse(gpanum<=3.49, gpa_level[3],
                   gpa_level[4])))

ds$gpaletter <- factor( classif, order=TRUE, levels=gpa_level)

#Comprobación
table(ds$gpaletter)
```

```
##
##    D    C    B    A
##  144 1536 1999  458
```

```
sum(table(ds$gpaletter))
```

```
## [1] 4137
```

```r
sum(table(ds$colgpa))
```

```
## [1] 4137
```

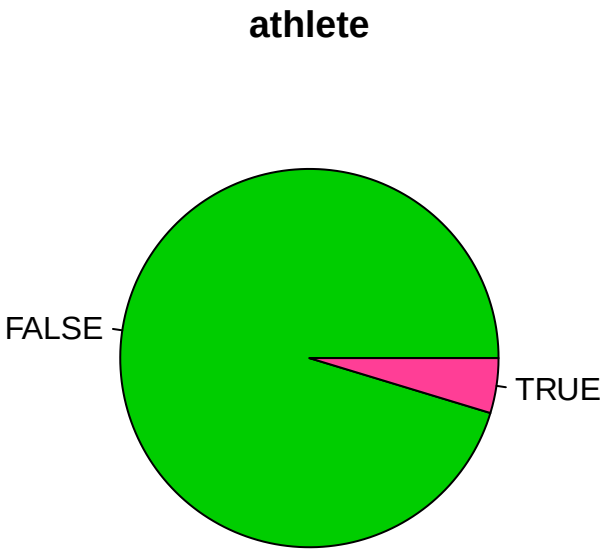
```r
#Reporting de canvis

  info <- report(info, row="*", "gpaletter: Nova variable que categoritza la nota numérica de colga en
```

7 Estudi descriptiu

7.1 Estudi descriptiu de les variables qualitatives

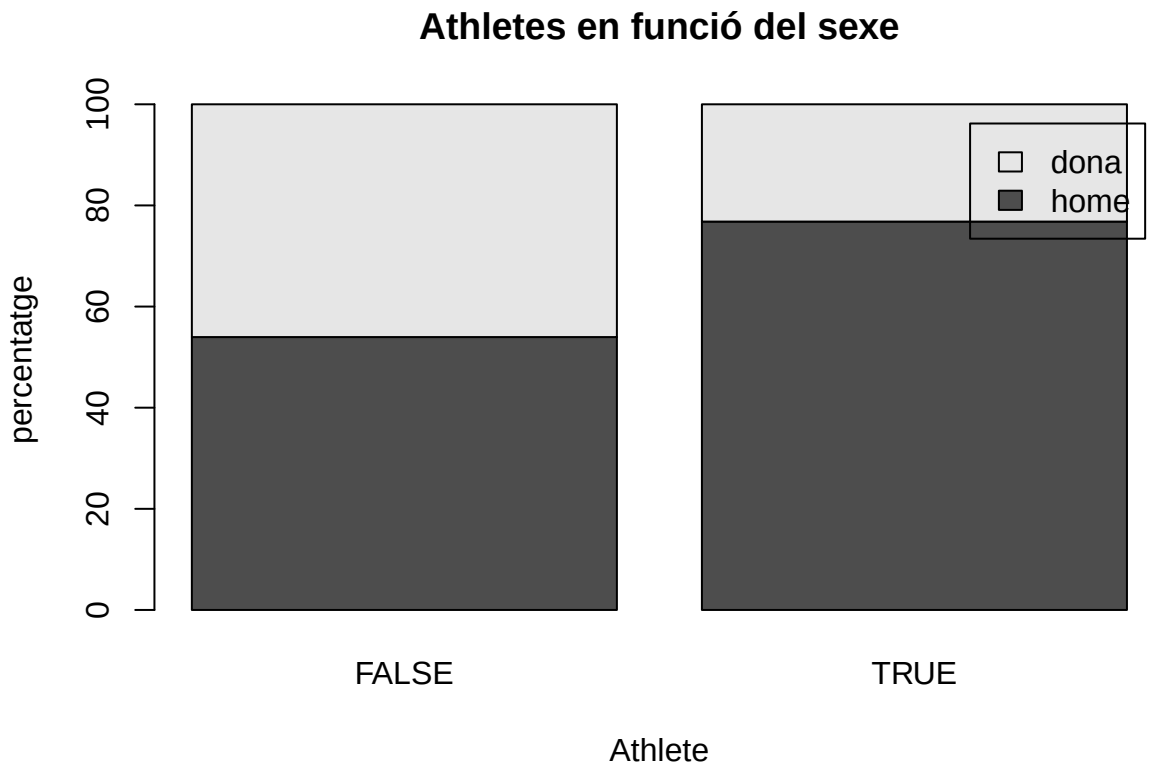
Representeu en un primer gràfic, la variable `athlete` en percentatge d'atletes i un segon gràfic, la variable `athlete` en funció del sexe on es mostri visualment si el percentatge de homes i dones canvia al ser atleta o no.

```r
# Representació gràfica pie plot
  pie(table(ds$athlete),
      main="athlete",
      col = c("green3", "violetred1")  )
```




```r
# Representació gràfica bar plot a %
tb1 <- table(ds$female, ds$athlete)
tb2 <- prop.table(tb1,margin = 2)*100
barplot(tb2,
        xlab="Athlete",
        ylab= "percentatge",
```

```
        main = "Athletes en funció del sexe",
        legend = c("home", "dona")  )
```



7.2 Estudi descriptiu de les variables quantitatives

Feu un estudi descriptiu de les variables quantitatives "sat", "tothrs", "hsize", "hsrank".

Per a això, prepareu una taula amb diverses mesures de tendència central i dispersió, robustes i no robustes. Presenteu els gràfics on es visualitzi la distribució dels valors de "sat" i "sat" en funció del sexe.

```
idx.numeric <- which( colnames(ds) %in% c("sat", "tothrs", "hsize", "hsrank") )
mean.n   <- as.vector(sapply( ds[,idx.numeric ],mean,na.rm=TRUE ) )
std.n    <- as.vector(sapply(ds[,idx.numeric ],sd, na.rm=TRUE))
median.n <- as.vector(sapply(ds[,idx.numeric], median, na.rm=TRUE))
mean.trim.0.05 <- as.vector(sapply( ds[,idx.numeric],mean, na.rm=TRUE, trim=0.05))
mean.winsor.0.05 <- as.vector(sapply( ds[,idx.numeric], winsor.mean, na.rm=TRUE,trim=0.05))

IQR.n <- as.vector(sapply(ds[,idx.numeric],IQR, na.rm=TRUE))
mad.n <- as.vector(sapply(ds[,idx.numeric],mad, na.rm=TRUE))

kable(data.frame(variables= names(ds)[idx.numeric],
                 Media = mean.n,
                 Mediana = median.n,
                 Media.recort.0.05= mean.trim.0.05,
                 Media.winsor.0.05= mean.winsor.0.05
                 ),
      digits=2, caption="Estimacions de Tendència Central")
```

Taula 3: Estimacions de Tendència Central

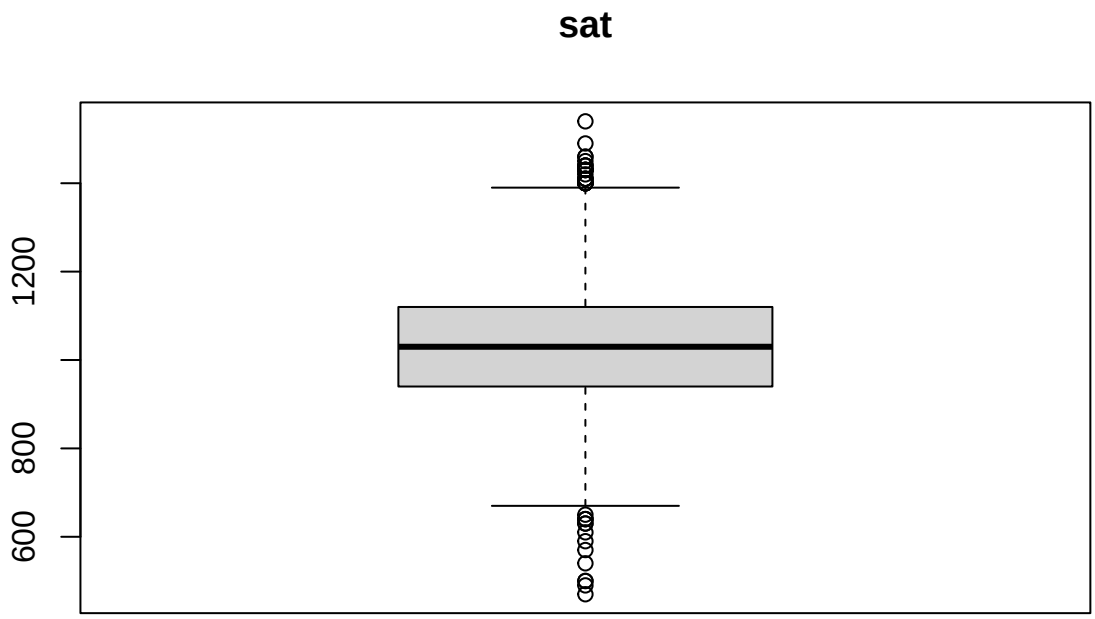
variables	Media	Mediana	Media.recort.0.05	Media.winsor.0.05
sat	1030.33	1030.00	1029.48	1030.53
tothrs	52.83	47.00	51.27	52.64
hsize	2.80	2.51	2.71	2.77
hsrank	52.83	30.00	43.99	48.78

Taula 4: Estimacions de Dispersió

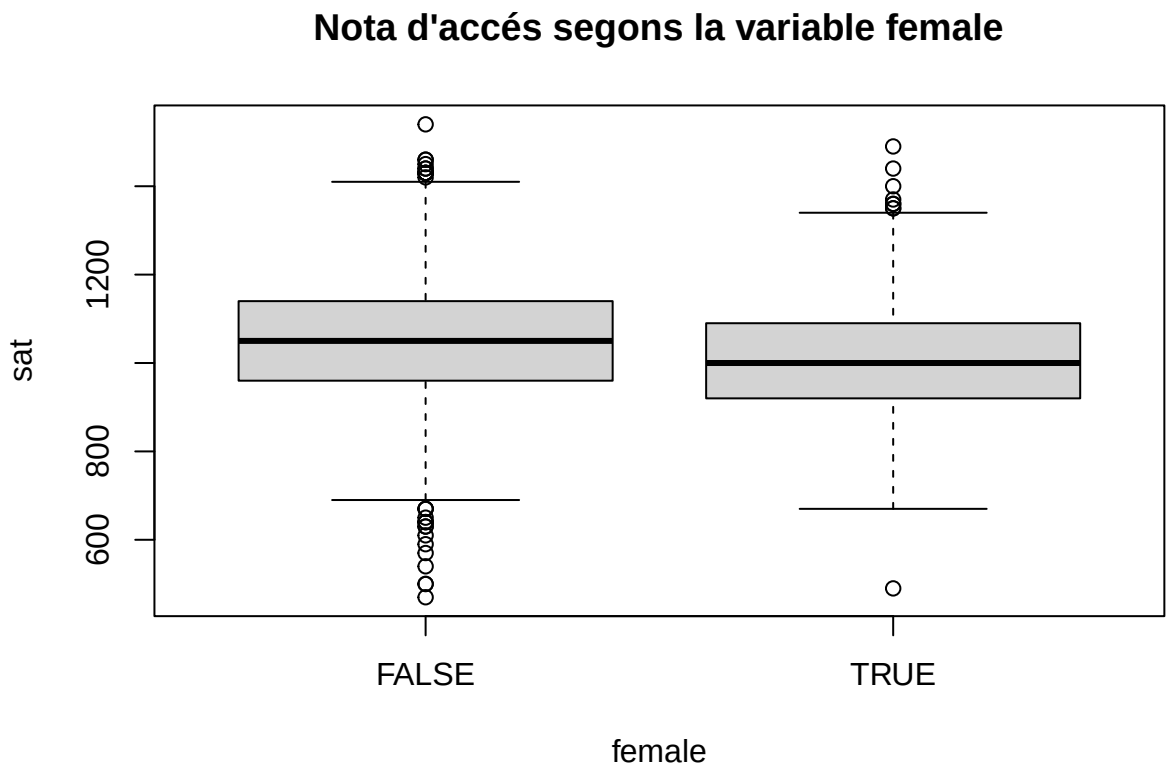
variables	Desv.Standard	IQR	MAD
sat	139.40	180.00	133.43
tothrs	35.33	63.00	45.96
hsize	1.74	2.03	1.42
hsrank	64.68	59.00	35.58

```
kable(data.frame(variables= names(ds)[idx.numeric],
                 Desv.Standard = std.n,
                 IQR = IQR.n,
                 MAD = mad.n
                 ),
      digits=2, caption="Estimacions de Dispersió")
```

```
#Gràfics
boxplot(ds$sat, main="sat")
```




```
boxplot(sat ~ female, data= ds, main="Nota d'accés segons la variable female")
```



8 Arxiu final

Un cop realitzat el preprocessament sobre l'arxiu, copieu el resultat de les dades a un fitxer anomenat **gpa_clean.csv**.

```
info <- report(info,row="",
                    message= paste0("n. row = ", nrow(ds),"; ",
                                    "n. col= ", ncol(ds), "; ",
                                    "n. var num. = ",  length(id.num),"; ",
                                    "n. var qualit. = ", length(id.factor)+1))


write.csv(ds, "gpa_clean.csv", row.names = FALSE)
```

9 Informe executiu

Està format per dos parts:

- Taula resum dels canvis realitzats al preprocessament.
- Breu explicació de les característiques estadístiques bàsiques de cada variable per separat.

9.1 Taula resum del preprocessament

Documentar de manera resumida, en forma de taula, els canvis introduïts a l'arxiu original durant el seu preprocessament. Cal explicar el detall del preprocés aplicat. Per exemple, no és suficient dir "s'ha normalitzat la variable hsize". En tot cas, s'hauria d'indicar si s'ha reemplaçat la coma pel punt decimal, o si s'han arrodonit decimals, etcètera i a quines observacions. Heu de ser específics, ja que l'informe ha de ser útil com a documentació dels canvis realitzats.

La primera i darrera fila de la taula ha d'indicar el nombre d'observacions, el nombre de variables quantitatives, el nombre de variables qualitatives i el total de variables a l'inici del preprocesament i al final, respectivament.

```
info <- info[1:nrow(info),]
info %>%
  kable( caption="Resum del preprocesament", row.names = FALSE)  %>%
  column_spec(2:3, width = "20em") %>%
  kable_styling( latex_options=c("striped", "repeat_"))
```

Taula 5: Resum del preprocesament

id	row	message
1		n. row = 4137; n. col= 10; n. var num. = 6; n. var qualit. = 4
1	65, 385, 720, 939, 1293, 1312, 2330, 2400, 2933, 3193, 3543	athlete: false -> FALSE
2	876	athlete: true -> TRUE
3	307, 754, 1230, 1858, 2213, 2374, 2376, 3042, 3173	black: Eliminat espais en blanc
4	1255, 1785, 2227, 2424, 2450, 2913, 3076, 3102, 3803, 3868	black: false -> FALSE
5	457, 595, 956, 2100, 3787, 3854	white: Eliminat espais en blanc
6	1929, 2922, 3536	white: false -> FALSE
7	461, 922, 1007, 1947, 2810, 2969, 3129, 4029, 4030	white: true -> TRUE
8	*	tothrs: Eliminat el text h i s'ha convertit a variable numèrica
9	53, 67, 155, 214, 371, 557, 565, 784, 842, 911, 948, 1399, 1566, 1723, 1956, 2024, 2293, 2304, 2361, 2382, 2603, 2689, 3832, 4003	corregim la coma pel punt decimal
10	188, 201, 313, 657, 876, 2489, 3438, 3441, 3445, 3537, 3753, 4091	hsperc: recalculem el valors amb hsrank/hsize
11	40, 100, 318, 343, 490, 500, 629, 846, 1053, 1172, 1226, 1238, 1319, 1450, 1605, 1866, 1888, 1937, 1975, 2035, 2108, 2184, 2530, 2536, 2691, 2721, 2728, 2879, 3149, 3196, 3495, 3496, 3523, 3546, 3651, 3660, 3758, 3798, 3943, 3998, 4015	41 registres amb NA a la variable colgpa
12	*	gpaletter: Nova variable que categoritza la nota numérica de colga en A, de 3.50 a 4.00; B, de 2.50 a 3.49; C, de 1.50 a 2.49; D, de 0 a 1.49
13		n. row = 4137; n. col= 11; n. var num. = 6; n. var qualit. = 5

9.2 Resum estadístic

A partir de la informació que s'ha obtingut en els apartats anteriors feu un breu comentari de cada variable destacant el més rellevant i característic. El resum no ha d'ocupar més d'una pàgina.

- **sat**: Variable numèrica. Representa la nota d'accés a la universitat. El valors de mitjana, mediana i medianes trimmed i winsor són molt similars, al voltant de 1030. Això indica que la distribució de les dades és pràcticament simètrica. Respecte a les estimacions de dispersió, tenen un valors similars desv. estandard i MAD, el IQR té un valor més alt de 180.
- **tothrs**: Variable numèrica. Representa el total d'hores cursades en el semestre. Els valors de mitjana estan al voltant de 52 hores, en canvi la mitjana és una mica menor, 47 hores. Això vol dir que hi ha estudiants amb valors molt alts que fan augmentar el valor de la mitjana respecte a la mediana. Els valors de dispersió varien. Així es té que la desv. estandard té el valor de 35.33 fins el IQR que val 63.
- **hsize**: Variable numèrica. Representa el nombre total d'estudiants a la cohort de graduats del batxillerat (en centenars). Els valors de les mitjanes són bastant similars entre 2.71 a 2.80. En canvi, el valor de mediana baixa una mica més a 2.51. És l'efecte d'alguns valors extrems. Respecte a la dispersió, es mou molt poc, entre un valor de MAD de 1.42 fins al IQR de 2.03.
- **hsrank**: Variable numèrica. Rànquing de l'estudiant donat per la nota mitjana del batxillerat de la cohort de graduats del batxillerat. Aquesta variable es la que presenta major diferència entre el valors de mitjana i mediana en comparació entre les altres variables, d'uns 23 punts. Això indica que la variable té alguns valors extrems. La mediana és de 30 punts mentre que la mitjana té un valor de 52.83. En qualsevol cas, com aquesta variable va associada a `hsize` és normal que si un institut té molts alumnes pugui tenir valors molt alts de `hsrank`. En canvi, en instituts petits la majoria dels valors de `hsize` quedaran en posicions baixes. Aquesta variabilitat també queda reflectida amb les estimacions de dispersió que van de 35.58 per a MAD fins a 64.68 per la desv. estandard.
- **hsperc**: Variable numèrica. Rànquing relatiu de l'estudiant. Com aquesta variable normalitza `hsrank` en funció de la mida de l'institut `hsize` no té aquesta diferència tan extrema entre els valors de mitjana i mediana i, els valors de estimacions de dispersió com `hsrank`. Encara que hi ha alguns valors extrems.
- **colgpa**: Variable numèrica. Representa la nota mitjana a final del primer semestre. Variable que té una mitjana i una mediana molt similars. Veien el boxplot s'observa que la distribució de valors és asimetrica amb una cua a l'esquerra des de cero fins al voltant de 2.4 que després baixa fins a 4.
- **athlete**: Variable categòrica binaria. Distribució molt desigual, la majoria dels estudiants (95%) no practiquen cap esport a la universitat.
- **female**: Variable categòrica binaria. Distribució molt similar amb un percentatge lleugerament superior d'estudiants homes (55,04%) respecte a dones (44,96%).
- **white**: Variable categòrica binaria. Distribució molt desigual. La majoria són white (92,55%)
- **black**: Variable categòrica binaria. Distribució molt desigual. Només un 5.55% dels alumnes són black. Si considerem la informació de les variables `white` i `black` conjuntament podem observar que hi ha uns 79 alumnes que no són ni black, ni white. Una proporció molt petita (0.02%).
- **gpaletter**: Variable categòrica amb quatre categories. La distribució de valors és D (3.48%), C(37.13%), B(48.32%) i A (11.07%). Així que la majoria dels alumnes aproven i una part important d'aquests amb unes notes altes.

10 Avaluació de l'activitat

- Seccions 1, 2 (10%)
- Seccions 3, 4 (20%)
- Secció 5 (10%)

- Secció 6 (10%)
- Seccions 7, 8 (20%)
- Secció 9 (20%)
- Qualitat de l'informe dinàmic (qualitat del codi, format, estructura del document, concisió i precisió en les respostes) (10%)

11 Referències

Quick-R

Cookbook for R

LaTex tables

Data Visualization with R