Estadística avançada
A1 Preprocessament de les dades

Noemi Lorente Torrelles

27 de octubre, 2018

Contents

1 Introducció

Les dades a tractar corresponen a una investigació mèdica orientada a estudiar la capacitat pulmonar de les persones en funció de si són fumadores o no.

A cada persona, se li va preguntar, a través d'un qüestionari, el seu gènere, edat, hàbits d'esport, la ciutat de residència, si era fumadora, i en cas que ho fos, quants cigarrets al dia de promig fumava i els anys que feia que fumava.

A més, es va mesurar el pes, l'alçada i la capacitat pulmonar a partir d'un test d'aire expulsat, des d'on es va prendre com a capacitat pulmonar la mesura FEF (forced expiratory flow), que és la velocitat de l'aire sortint del pulmó durant la porció central d'una espiració forçada. Es mesura en litres/segon.

L'arxiu és denomina Fumadores_raw.csv, conté 300 registres i 9 variables.

Aquestes variables són: Sex, Sport, Years, Cig, PC, City, Weight, Age, Height.

2 Preprocessament de dades

L'objectiu concret d'aquesta activitat és preparar el fitxer per a la seva posterior anàlisi.

Per a guiar l'activitat us suggerim quin tipus de preprocessament cal fer per a resoldre satisfactòriament aquesta activitat.

2.1 Criteris a seguir en el preprocessament de dades

Criteris a seguir en el preprocessament de dades:

- El punt (.) és el separador decimal de qualsevol variable numèrica.
- Els valors de la variable PC s'estandaritzen en tres xifres decimals.
- Els noms de les ciutats s'estandaritzen amb la primera lletra de cada paraula en majúscules i la resta en minúscules. Per exemple: Barcelona, Tarragona, Lleida, Girona, . . .
- Els noms de gènere s'estandaritzen com M i F.
- Els noms dels hàbits d'esport s'estandaritzen amb una lletra d'acord amb l'equivalència: 1 equival a N (None), 2 equival a S (Sometimes), 3 equival a R (Regularly) i 4 equival a E (Every day).
- La inconsistència entre les variables Years vs Cig es produeix quan individu no fumador (Years = 0) respon que fuma algun cigarret (Cig > 0) o també, quan un individu fumador (Years > 0) respon que no fuma cap cigarret (Cig = 0). En aquests casos, el criteri de correcció és assignar a les dues variables, Years i Cig, el valor de zero.
- En cas de trobar algun valor de la variable pes atípic és degut a que va ser registrat en grams. Cal que expresseu el pes en quilograms.

2.2 Preprocessaments a realitzar

Els preprocessaments a realitzar són els següents:

2.2.1 Carregar fitxer de dades i breu descripció

Carregar el fitxer de dades en R i fer una breu descripció de l'arxiu on s'indiqui {el nombre de registres, el nombre de variables i els noms de les variables.

Abans de carregar el fitxer, he visualitzat el seu contingut amb *gedit* i he pogut comprobar que s'utilitza la coma "," com a separador de camps. Així m'asseguro que la lectura del fitxer es realitza de forma correcta.

```
# carrego el fitxer amb read.table, separador=',', decimal='.'

capacitat <- read.table("Fumadores_raw.csv", header=TRUE, sep=",", na.strings="NA",
                        dec=".", strip.white=TRUE, stringsAsFactors = FALSE)
```

Amb la funció **str** puc veure l'estructura interna del data frame *capacitat*. Veig que té 300 observacions, 9 variables i es mostra el nom de les variables.

```r
# la funció str mostra l'estructura interna del data frame capacitat

str(capacitat)
```

```
## 'data.frame':    300 obs. of  9 variables:
##  $ Sex   : chr  "M   " "F" "M" "M   " ...
##  $ Sport : int  1 1 4 2 1 4 1 1 1 1 ...
##  $ Years : int  25 18 0 25 0 0 33 0 0 5 ...
##  $ Cig   : int  10 32 0 14 0 0 15 0 1 12 ...
##  $ PC    : chr  "2.57917010238911" "1.5570436932541" NA "2.76216877363087" ...
##  $ City  : chr  "Barcelona" "  Terrassa" "  La Bisbal" "Blanes   " ...
##  $ Weight: int  65 65 69 70 72 64 69 71 72000 73 ...
##  $ Age   : int  49 35 38 55 55 42 55 44 45 35 ...
##  $ Height: int  171 166 175 176 178 165 175 177 178 179 ...
```

2.2.2 Tipus de variables estadístiques

Indicar el tipus de variable estadística que pertocaria a cada variable. És a dir, quines variables haurien de ser: qualitatives nominals, quantitatives discretes i quantitatives contínues.

Variable	Tipus variable
Sex:	qualitativa nominal
Sport:	qualitativa nominal
Years:	quantitativa discreta
Cig:	quantitativa discreta
PC:	quantitativa continua
City:	qualitativa nominal
Weight:	quantitativa continua
Age:	quantitativa discreta
Height:	quantitativa continua

2.2.3 Conversió de tipus de variables

En cas que R no assigni el tipus apropiat a alguna variable, cal realitzar la conversió necessària per a que el tipus final de cada variable sigui adequat. Per exemple, revisar si hi ha variables que haurien de ser de tipus "factor". I pel que fa a les variables numèriques, revisar si aquestes han estat degudament interpretades com a tal. Si cal fer la conversió a numèric, prèviament detecteu si hi ha possibles errors en el separador decimal i si és el cas, corregiu-los abans de fer la conversió.

- Les variables **Sex**, **Sport** i **City** han de ser factor:

```r
# Canvio a variable factor la variable Sex
capacitat$Sex <- as.factor(capacitat$Sex)
```

```r
# Canvio a variable factor la variable Sport
capacitat$Sport <- as.factor(capacitat$Sport)
```

```
# Canvio a variable factor la variable City
capacitat$City <- as.factor(capacitat$City)
```

- La variable **PC** ha de ser numèrica. Abans de realitzar el canvi de tipus de variable, cal assegurar que tots els valors fan anar el mateix símbol com a separador decimal, el punt (.) Com aquesta variable conté alguna coma (,) de separador decimal, cal cercar la coma i substituir-ho per punt. Utilitzo la funció **str_replace** del package *stringr*.

```
# Reemplaço la "," per "." en la variable PC
library(stringr)
capacitat$PC<-str_replace(capacitat$PC, ",", ".")
```

```
# Canvio a variable numèrica la variable PC
capacitat$PC <- as.numeric(capacitat$PC)
```

Per conèixer el tipus de variable utilitzo la funció **class**. A continuació, mostro el tipus de variable de cada variable. Per mostrar els tipus de les variables en una taula, utilitzo la funció **kable** .

```
# Recupero el tipus de variable
tipus <- sapply(capacitat,class)
kable(data.frame(Variable=names(tipus),Classe=as.vector(tipus)), align='l',
      caption="Tipus de les variables")
```

Table 2: Tipus de les variables

Variable	Classe
Sex	factor
Sport	factor
Years	integer
Cig	integer
PC	numeric
City	factor
Weight	integer
Age	integer
Height	integer

2.2.4 Normalitzar/Estandarditzar variables quantitatives

Normalitzar/Estandarditzar variables quantitatives, seguint els criteris de preprocessament de dades.

De moment, no es requerix cap normalització z-score de les variables quantitatives per a l'anàlisi estadística, ja que no es demana comparar variables entre si.

Es poden diferenciar tres tipus de casos d'errors sintàctics:

1. Confusió en l'ús de diferents símbols com a separadors decimals.

En la variable **PC**, he unificat el separador decimal a punt (.) en l'apartat *2.2.3 Conversió de tipus de variables.*

2. Estandarditzar les variables a les mateixes unitats.

Els valors de la variable **PC** cal estandaritzar-los en tres xifres decimals. Utilitzo la funció **round**.

```
# Arrodonim a 3 xifres decimals la variable PC
capacitat$PC <- round(capacitat$PC, digits=3)
```

3. Estandarditzar les variables de data i temps.

El dataset no conté cap variable tipus Date ni POSIXct.

2.2.5 Normalitzar/Estandarditzar variables qualitatives

Normalitzar/Estandarditzar variables qualitatives, seguint els criteris de preprocessament de dades.

Es poden diferenciar tres tipus de casos d'errors sintàctics:

1. Treure espais abans i després del text o treure caràcters especials com el retorn de carro o tabulació.

Les variables **Sex** i **City** contenen espais abans i després del text. Utilitzo la funció **trimws** del package *stringr*.

```
# Elimino espais de la variable Sex
capacitat$Sex <- trimws(capacitat$Sex)

# Elimino espais de la variable City
capacitat$City <- trimws(capacitat$City)
```

2. Estandarditzar les variables a minúscules o majúscules.

- En la variable **Sex**, els noms de gènere s'estandaritzen com M i F. Cal passar-ho a majúscules. Utilitzo la funció **toupper** del package *stringr*.

```
# En la variable Sex, tot majúscules
capacitat$Sex <- toupper(capacitat$Sex)
```

- En la variable **City**, els noms de les ciutats s'estandaritzen amb la primera lletra de cada paraula en majúscules i la resta en minúscules. Per exemple: Barcelona, Tarragona, Lleida, Girona,... Utilitzo la funció **stri_trans_general** del package *stringr*.

```
# En la variable City, primera lletra de cada paraula en majúscules
# i la resta en minúscules
require(stringi)
```

```
## Loading required package: stringi
```

```
capacitat$City <- stri_trans_general(capacitat$City, id = "Title")
```

3. Eliminar accents.

La variable City pot contenir algún accent. Utilitzo la funció **stri_encode** del package *stringr*.

```
# Elimino els possibles accents de la variable City
capacitat$City <- iconv(capacitat$City, to="ASCII//TRANSLIT")
```

4. Revisió de categories errònies per errors sintàctics. Es pot detectar si es revisa la taula de freqüències de cada categoria. Utilitzo la funció **table** per obtenir la taula de freqüències.

- En la variable **Sex** només hi ha dos valors possibles: F (137 observacions) i M (163 observacions).

```
# Taula de freqüències de la variable Sex
kable(
      table(capacitat$Sex, dnn=c("Sex")), align='c',
      caption="Distribució de freqüències de la variable Sex")
```

Table 3: Distribució de freqüències de la variable Sex

Sex	Freq
F	137
M	163

- En la variable **City** hi ha 20 valors diferents i són correctes.

```
# Taula de freqüències de la variable City
kable(
      table(capacitat$City, dnn=c("City")), align='l',
      caption="Distribució de freqüències de la variable City")
```

Table 4: Distribució de freqüències de la variable City

City	Freq
Alcanar	12
Barcelona	102
Blanes	11
Cadaques	2
Cardedeu	3
Cardona	6
Girona	10
La Bisbal	7
Lleida	13
Montgat	6
Pineda De Mar	11
Puigcerda	3
Ripoll	6
Sant Boi De Llobregat	12
Sitges	13
Solsona	5
Tarragona	14
Terrassa	42
Tossa	7
Valls	15

5. Altres estandaritzacions del model de negoci

- En la varible **Sport**, els noms dels hàbits d'esport s'estandaritzen amb una lletra d'acord amb l'equivalència: 1 equival a N (None), 2 equival a S (Sometimes), 3 equival a R (Regularly) i 4 equival a E (Every day).

```
# En la variable Sport, creo factor intercanviant els nivells 1, 2, 3, 4 per N, S, R i E
capacitat$Sport <- factor(capacitat$Sport, levels=c(1,2,3,4),
                          labels=c("N", "S", "R", "E"))
```

```
# Taula de freqüències de la variable Sport
kable(table(capacitat$Sport, dnn=c("Sport")), align='l',
      caption="Distribució de freqüències de la variable Sport")
```

Table 5: Distribució de freqüències de la variable Sport

Sport	Freq
N	127
S	83
R	48
E	42

2.2.6 Revisió d'inconsistències entre variables

Revisar possibles inconsistències entre variables: Years vs Cig, seguint els criteris de preprocessament de dades.

Les inconsistències són errors que es poden haver produït en el moment de la introducció de la informació i es poden reconèixer fàcilment.

Concretament, cal revisar la següent inconsistència entre les variables **Years** i **Cig**: es produeix quan individu no fumador (Years = 0) respon que fuma algun cigarret (Cig > 0) o també, quan un individu fumador (Years > 0) respon que no fuma cap cigarret (Cig = 0). En aquests casos, el criteri de correcció és assignar a les dues variables, Years i Cig, el valor de zero.

```
# A la variable Cig li assigno
# el valor 0 en cas que es tracti d'un no fumador (Years = 0)
# que respon que fuma algun cigarret (Cig > 0). Hi ha 21 casos d'inconsistència.
length( which(capacitat$Years== 0 & capacitat$Cig>0) )
```

```
## [1] 21
```

```
capacitat$Cig[capacitat$Years== 0 & capacitat$Cig>0] <- 0
length( which(capacitat$Years== 0 & capacitat$Cig>0) )
```

```
## [1] 0
```

```
# A la variable Years li assigno
# el valor 0 en cas que es tracti d'un fumador (Years > 0)
# que respon que no fuma cap cigarret (Cig = 0). Hi ha 21 casos d'inconsistència.
length( which(capacitat$Years> 0 & capacitat$Cig==0) )
```

```
## [1] 21
```

```
capacitat$Years[capacitat$Years> 0 & capacitat$Cig==0] <- 0
length( which(capacitat$Years> 0 & capacitat$Cig==0) )
```

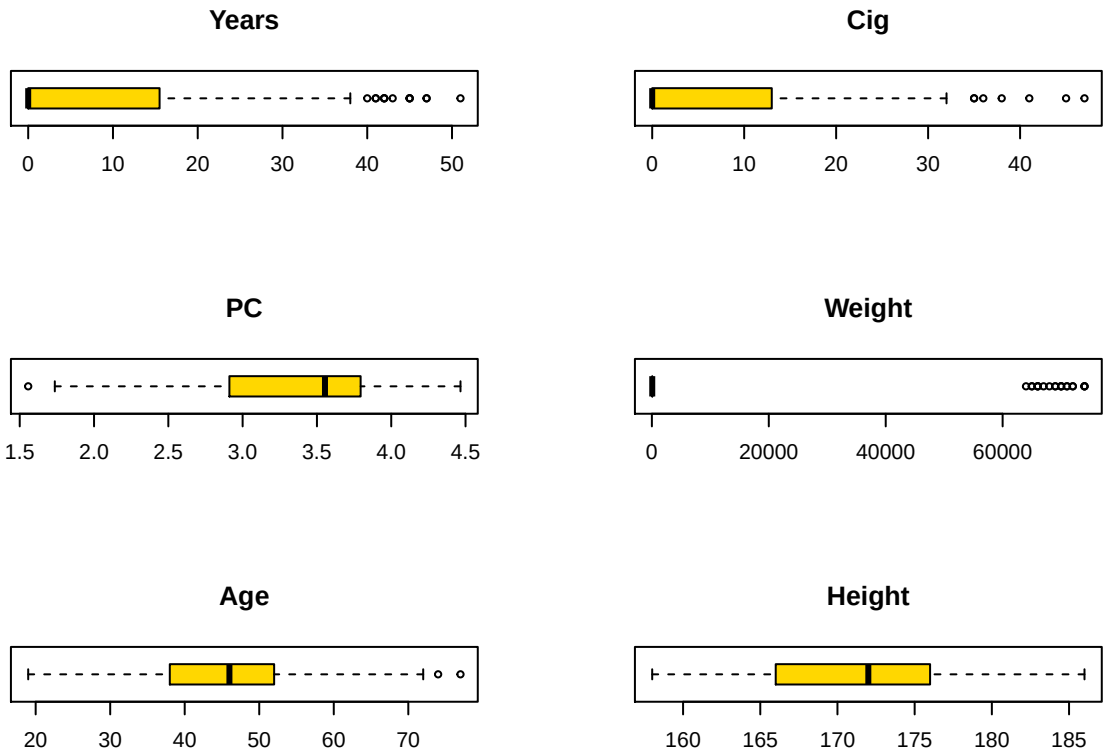
```
## [1] 0
```

2.2.7 Valors atípics en variables quantitatives

Buscar valors atípics en les variables quantitatives.

2.2.7.1 Presentar un boxplot per cada variable quantitativa.

```
# Genero boxplot de Years, Cig, PC, Weight, Age, Height
par(mfrow=c(3,2))
boxplot(capacitat$Years, main="Years", horizontal=TRUE, col="gold")
boxplot(capacitat$Cig, main="Cig", horizontal=TRUE, col="gold")
boxplot(capacitat$PC, main="PC", horizontal=TRUE, col="gold")
boxplot(capacitat$Weight, main="Weight", horizontal=TRUE, col="gold")
boxplot(capacitat$Age, main="Age", horizontal=TRUE, col="gold")
boxplot(capacitat$Height, main="Height", horizontal=TRUE, col="gold")
```



2.2.7.2 Si observeu algun valor anòmal en alguna variable, considereu la seva transformació/correcció.

Observo que la variable **Weight** conté valors atípics.

Per saber quins valors atípics té la variable **Weight** utilizo el valor *out* de *boxplot.stats*.

```
# Per saber quin son els valors atípics de la variable Weight
boxplot.stats(capacitat$Weight)$out
```

```
##  [1] 72000 69000 72000 70000 74000 66000 70000 65000 74000 71000 68000
## [12] 74000 70000 66000 67000 65000 64000 69000 66000 71000
```

Hi ha 20 valors atípics i comprovo que es corresponen amb els 20 valors majors de 1000. Visualitzo les dades amb la funció **head**, així puc observar la diferencia que es produirà amb la correcció.

```
# En la variable Weight hi ha valors > 1000 perque es van registrar en grams
length( which(capacitat$Weight > 1000) )
```

```
## [1] 20
```

```
head(capacitat$Weight,20)
```

```
##  [1]    65    65    69    70    72    64    69    71 72000    73 69000
## [12]    65 72000 70000    65    60    66    66    70    63
```

En els criteris a seguir en el preprocessament de les dadesm, s'indica que en cas de trobar algun valor de la variable **Weight** atípic és degut a que va ser registrat en grams. Per tant, cal transformar el pes en quilograms.

```
# Per passar-ho a kilograms, dividim el valor entre 1000
capacitat$Weight[capacitat$Weight > 1000] <-
  capacitat$Weight[capacitat$Weight > 1000] / 1000

head(capacitat$Weight,20)
```

```
##  [1] 65 65 69 70 72 64 69 71 72 73 69 65 72 70 65 60 66 66 70 63
```

```
length( which(capacitat$Weight > 1000) )
```

```
## [1] 0
```

2.2.7.3 Realitzar un quadre amb les estimacions robustes i no robustes de tendència central i dispersió per a cada variable quantitativa.

El subcamp de l'estadística denominat estadística robusta (Huber,1981; Leroy, 1987; Maronna i altres, 2006) desenvolupa estimadors i models estadístics alternatius a l'estadística clàssica, però amb la condició que els estimadors siguin resistents (robustos) a l'efecte dels valors atípics i els models estadístics puguin admetre certes desviacions en les condicions d'aplicabilitat.

Selecciono les variables quantitatives del data frame *capacitat* i ho guardo en la variable *VbleQuant*.

```
# Selecciono les variables quantitatives del data frame capacitat
VbleQuant <- capacitat[,c("Years", "Cig", "PC", "Weight", "Age", "Height")]
```

Els estimadors robustos de tendència central són els següents:

- *mitjana*: tot i que té molt bones propietats, la seva estimació està molt influïda pels valors atípics, per tant, no és robusta però afegeixo el seu càlcul a la taula per poder comparar amb la resta d'estimadors robustos. Utilitzo la funció **mean**.
- *mediana*: és un dels estimadors robustos més coneguts per a mesurar la tendència central de les dades. S'utilitza la funció **median**.
- *mitjana retallada (trimmed mean)*: Es basa a eliminar el mateix percentatge, k %, dels valors extrems, i calcular la mitjana aritmètica de la resta de valors. S'utilitza la funció **mean**, indicant el percentatge a eliminar en l'argument *trim*. En aquest cas, 5%.
- *mitjana winsoritzada (winsorized mean)*: Variant de la mitjana retallada. Els valors extrems eliminats són substituïts per altres valors. El cas més senzill és substituir pel valor més petit (o més gran) que ha quedat. S'utilitza la funció **winsor.mean** del package *psych*, indicant el percentatge a eliminar en l'argument *trim*. En aquest cas, 5%.

```
# Calculo els estimadors robustos de tendència central
require(psych)
```

```
## Loading required package: psych
```

```
mitjana <- as.vector(sapply(VbleQuant, mean,
                            na.rm = TRUE))
mediana <- as.vector(sapply(VbleQuant, median,
                            na.rm = TRUE))
mitjana.retallada <- as.vector(sapply(VbleQuant, mean,
                                      trim = 0.05, na.rm = TRUE ))
mitjana.winsor <-  as.vector(sapply(VbleQuant, winsor.mean,
                                    trim = 0.05, na.rm = TRUE))
```

Per a crear la taula amb els valors dels estimadors robustos de tendència central per a cada una de les variables quantitatives, també utilitzo la funció **kable**.

```
# Creo la taula amb els valors dels estimadors robustos de tendència central
# per a cada variable quantitativa

kable(
  data.frame(
    variables = names(VbleQuant),
    Mitjana = mitjana,
    Mediana = mediana,
    Mitjana_Retallada = mitjana.retallada,
    Mitjana_Winsortitzada = mitjana.winsor
  ),
  digits = 2,
  align='c',
  caption = "Estimacions robustes de tendència central"
)
```

Table 6: Estimacions robustes de tendència central

variables	Mitjana	Mediana	Mitjana_Retallada	Mitjana_Winsortitzada
Years	8.46	0.00	7.01	8.11
Cig	7.27	0.00	6.11	7.05
PC	3.33	3.55	3.36	3.33
Weight	67.72	68.00	67.73	67.71
Age	45.59	46.00	45.49	45.54
Height	171.44	172.00	171.47	171.43

La mediana de **Years** i **Cig** és zero perquè més de la meitat dels valors de les variables són zero. Observo que la Mitjana_Winsortitzada és bastant similar a la mitjana normal en totes les variables, excepte per a la variable **Years**.

Els estimadors robustos de dispersió són els següents:

- La *desviació estàndard*: és l'estimador de dispersió més habitual, però no és un estimador robust. Afegeixo el seu càlcul a la taula per poder comparar amb la resta d'estimadors robustos. Utilitzo la funció **sd**.
- El *rang interquartílic (RIC)*: diferència entre el quartil 3 (percentil 75 %) i el quartil 1 (percentil 25 %) de les dades. És la grandària de la caixa del gràfic box plot. Utilitzo la funció **IRQ**.
- La *desviació absoluta respecte de la mediana (DAM)*: és un estimador anàleg de la desviació estàndard però utilitza la mediana en lloc de la mitjana aritmètica com a valor central de les dades: es calcula el valor de la mediana de la diferència dels valors absoluts entre les dades i la seva mediana. Utilitzo la funció **MAD**.

$MAD(i) = mediana(|yi - mediana(y)|)$

```
# Calculo els estimadors robustos de dispersió
desviacio <- as.vector(sapply(VbleQuant, sd,
                              na.rm = TRUE))
rangInterQ <- as.vector(sapply(VbleQuant, IQR,
                               na.rm = TRUE))
desviacioAbs <- as.vector(sapply(VbleQuant, mad,
                                 na.rm = TRUE))
```

```
# Creo la taula amb els valors dels estimadors robustos de dispersió
# per a cada variable quantitativa
kable(
  data.frame(
    variables = names(VbleQuant),
    Desviacio_Estandard = desviacio,
    Rang_Interquartilic = rangInterQ,
    Desviacio_Absoluta_MAD = desviacioAbs
  ),
  digits = 2,
  align='c',
  caption = "Estimacions robustes de dispersió"
)
```

Table 7: Estimacions robustes de dispersió

variables	Desviacio_Estandard	Rang_Interquartilic	Desviacio_Absoluta_MAD
Years	12.54	15.25	0.00
Cig	10.41	13.00	0.00
PC	0.63	0.88	0.54
Weight	3.83	6.00	4.45
Age	10.63	14.00	10.38
Height	5.74	10.00	7.41

2.2.8 Valors perduts.

Revisar els valors perduts.

2.2.8.1 Buscar quines variables i registres tenen valors perduts.

La variable **Sex** no té valors perduts.

```
# Comprovo si la variable Sex té valors perduts
length( which(is.na(capacitat$Sex)))
```

```
## [1] 0
```

La variable **Sport** no té valors perduts.

```
# Comprovo si la variable Sport té valors perduts
length( which(is.na(capacitat$Sport)))
```

```
## [1] 0
```

La variable **Years** no té valors perduts.

```
# Comprovo si la variable Years té valors perduts
length( which(is.na(capacitat$Years)))
```

```
## [1] 0
```

La variable **Cig** no té valors perduts.

```
# Comprovo si la variable Cig té valors perduts
length( which(is.na(capacitat$Cig)))
```

```
## [1] 0
```

La variable **PC** té 2 valors perduts.

```
# Comprovo si la variable PC té valors perduts
length( which(is.na(capacitat$PC)))
```

```
## [1] 2
```

La variable **City** no té valors perduts.

```
# Comprovo si la variable City té valors perduts
length( which(is.na(capacitat$City)))
```

```
## [1] 0
```

La variable **Weight** no té valors perduts.

```
# Comprovo si la variable Weight té valors perduts
length( which(is.na(capacitat$Weight)))
```

```
## [1] 0
```

La variable **Age** no té valors perduts.

```
# Comprovo si la variable Age té valors perduts
length( which(is.na(capacitat$Age)))
```

```
## [1] 0
```

La variable **Height** no té valors perduts.

```
# Comprovo si la variable Height té valors perduts
length( which(is.na(capacitat$Height)))
```

```
## [1] 0
```

2.2.8.2 Imputar els valors a partir dels k-veïns més propers usant la distància de Gower amb la informació de totes les variables.

El paquet *VIM* conté la funció **kNN** que utilitza la distància de *Gower* per a la imputació a partir del k-veïns més propers.

Els principals arguments de la funció kNN són:

- *dist_var*: un vector de les variables a emprar per a calcular les distancies. Com cal emprar la informació de totes les variables, no es necessari utilitzar aquest argument.
- *weights*: un vector numèric que conté el pes de cada variable. Tampoc aplica en aquest cas.
- *numFun*: la funció d'aggregació dels k-veïns propers per a les variables quantitatives. Per defecte és la mediana.
- *catFun*: la funció d'aggregació dels k-veïns propers per a les variables quantitatives. Per defecte és la funció maxCat, que retorna el valor més freqüent, i si aquest no és únic, aleshores un valor aleatori.
- *imp_var*: si es volen crear variables booleanes que indiquen l'estat de la imputació de variables. Com no vull que es generin aquestes variables, indico *imp_var = FALSE*.

```
# imputació basada en els k veïns més propers («kNN-imputation»)
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```
capacitat <- kNN(capacitat, k=5, numFun = median, catFun = maxCat, imp_var = FALSE)
```

Despres del procés d'imputació, observo que la variable **PC** ja no té valors perduts.

```
length( which(is.na(capacitat$PC)))
```

```
## [1] 0
```

2.2.9 Estudi descriptiu

Fer un breu estudi descriptiu de les dades una vegada depurades.

Per obtenir una estadística descriptiva simple de cada variable utilitzo la funció **summary**.

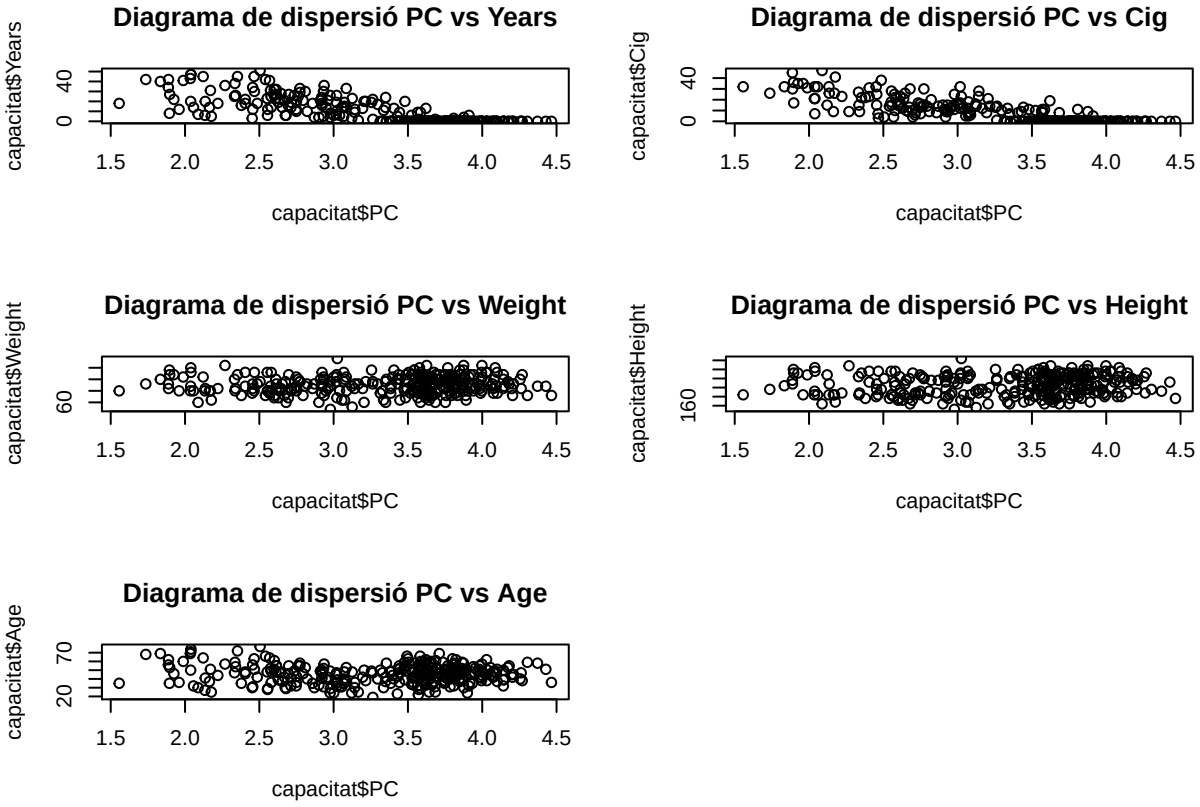
Per a variables quantitatives es mostren algunes mesures de tendència central, com la mitjana, mediana, mínim, màxim i el primer i tercer quartil.

```
# la funció summary retorna un resum de les mesures de tendència central
# per a variables quantitatives



kable(summary(capacitat[,c("Years", "Cig", "PC", "Weight", "Age", "Height")]),
      digits=2,
      align='c',
      caption="Estadística descriptiva de variables quantitatives")
```

Table 8: Estadística descriptiva de variables quantitatives

Years	Cig	PC	Weight	Age	Height
Min. : 0.000	Min. : 0.000	Min. :1.557	Min. :57.00	Min. :19.00	Min. :158.0
1st Qu.: 0.000	1st Qu.: 0.000	1st Qu.:2.909	1st Qu.:65.00	1st Qu.:38.00	1st Qu.:166.0
Median : 0.000	Median : 0.000	Median :3.554	Median :68.00	Median :46.00	Median :172.0
Mean : 8.463	Mean : 7.273	Mean :3.331	Mean :67.72	Mean :45.59	Mean :171.4
3rd Qu.:15.250	3rd Qu.:13.000	3rd Qu.:3.793	3rd Qu.:71.00	3rd Qu.:52.00	3rd Qu.:176.0
Max. :51.000	Max. :47.000	Max. :4.466	Max. :79.00	Max. :77.00	Max. :186.0

Observo que no hi ha relació lineal entre la variable PC i la resta de variables.



Per a variables qualitatives de tipus factor, amb la funció **summary** es mostren les freqüencies absolutes d'alguns dels valors o bé el nombre de registres, el tipus de variable i el nombre de valors perduts.

```
# la funció summary retorna el nombre de registres i el tipus de classe
# per a variables qualitatives


kable(summary(capacitat[,c( "Sex", "Sport", "City")]),
      digits=2,
      align='c',
      caption="Estadística descriptiva de variables qualitatives")
```

Table 9: Estadística descriptiva de variables qualitatives

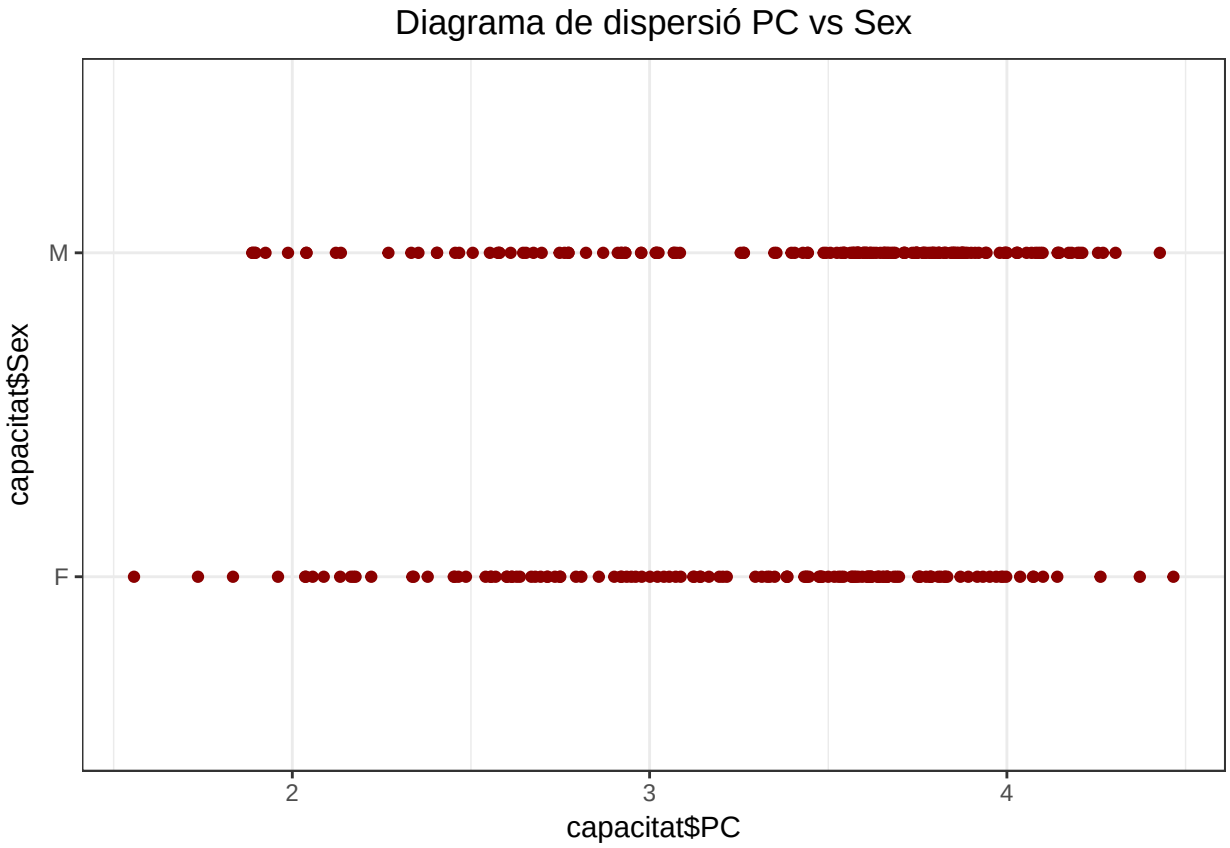
Sex	Sport	City
Length:300	N:127	Length:300
Class :character	S: 83	Class :character
Mode :character	R: 48	Mode :character
NA	E: 42	NA

També visualitzem els diagrames de dispersió entre la variable **PC** i les variables qualitatives.

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```



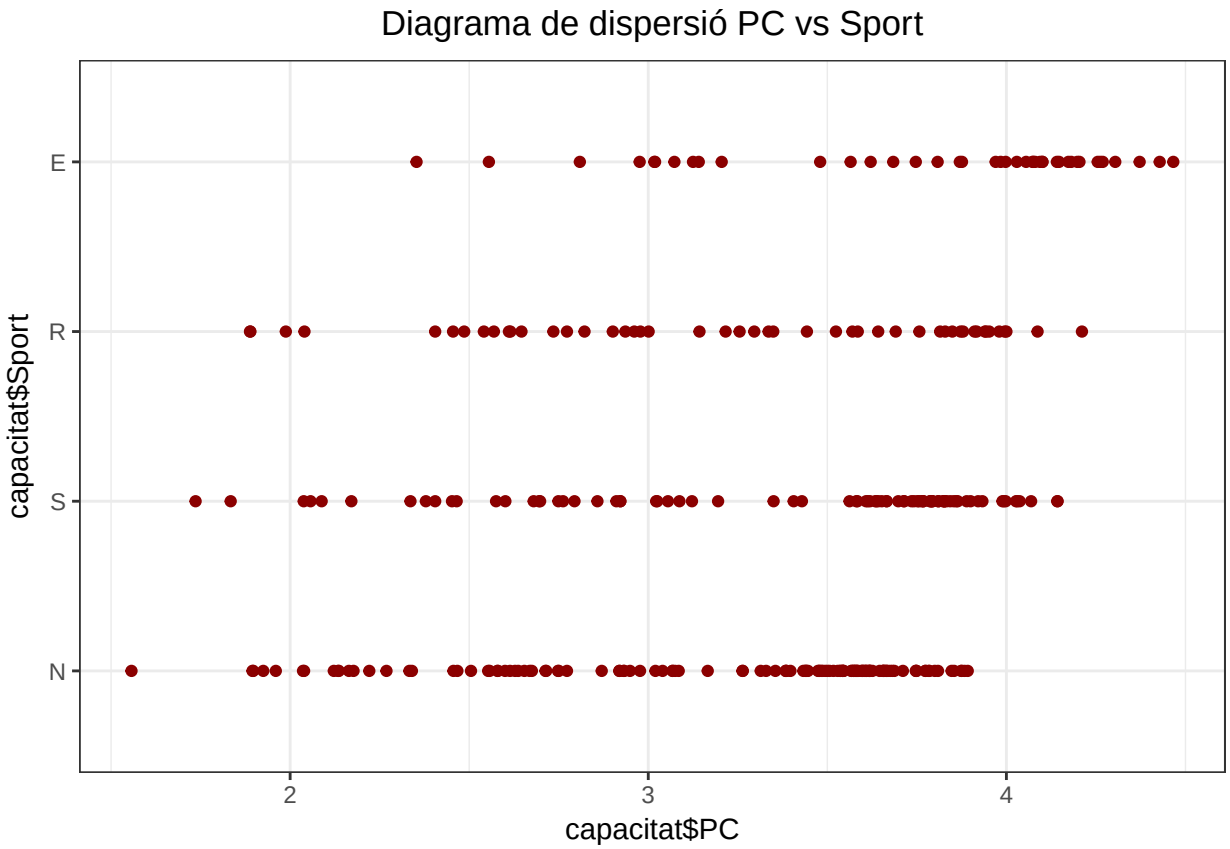
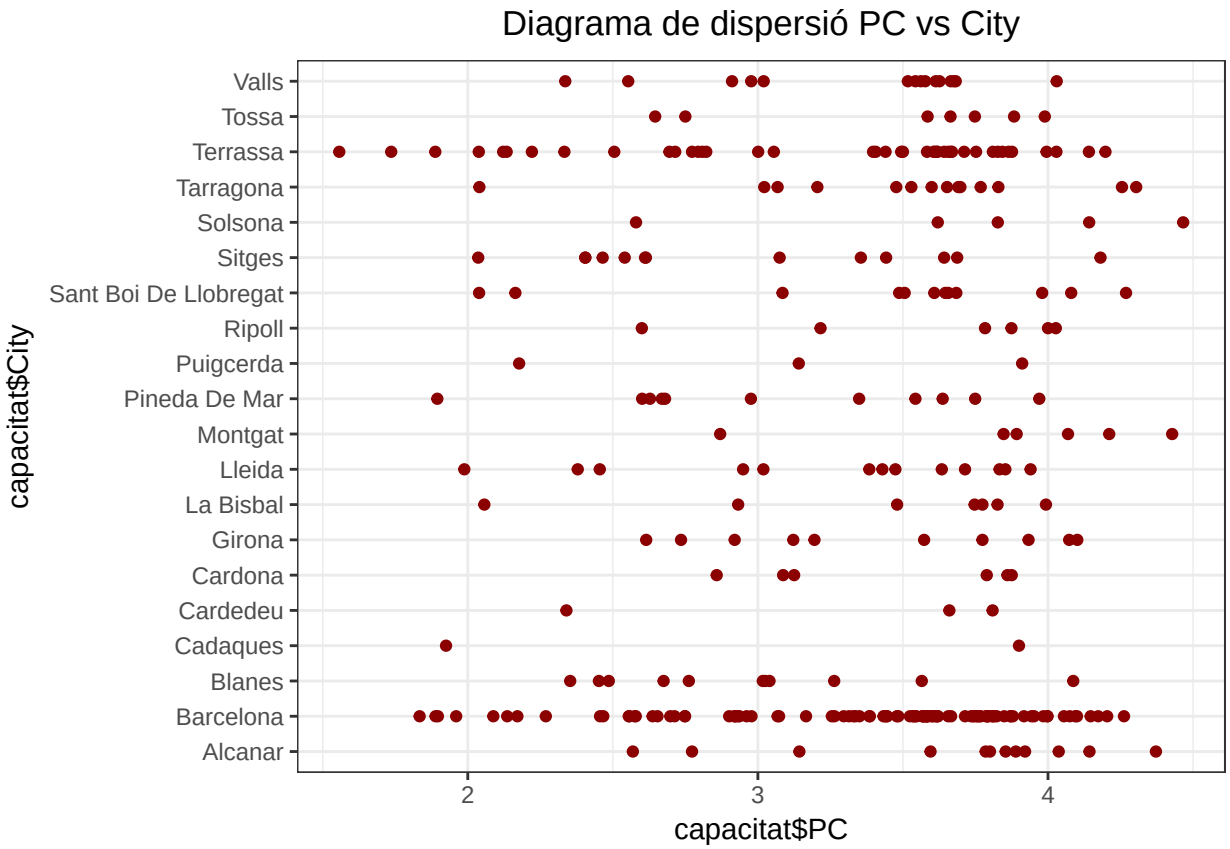


Diagrama de dispersió PC vs Sport
capacitat$Sport
E
R
S
N
2
3
4
capacitat$PC



2.2.10 Creació de l'arxiu de dades corregit

Finalment, crear l'arxiu de dades corregit.

Guardo el conjunt de dades processat *capacitat* amb el nom *capacitat.csv*. Utilitzo la funció **write.csv**.

```
# Estudi descriptiu amb un summary
write.csv(capacitat, file="capacitat.csv", row.names = FALSE)
```

3 Referències

- Rmarkdown cheat sheet https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf
- Rmarkdown: The Definitive Guide https://bookdown.org/yihui/rmarkdown/pdf-document.html
- RDocumentation kNN https://www.rdocumentation.org/packages/VIM/versions/4.7.0/topics/kNN
- RDocumentation write.table https://www.rdocumentation.org/packages/utils/versions/3.5.1/topics/write.table
- RDocumentation kable https://www.rdocumentation.org/packages/knitr/versions/1.20/topics/kable