PAC 1

Carlos A. García

March 19, 2019

1. Carregueu l'arxiu de dades en R i feu una breu descripció del mateix, on s'indiqui el nombre de registres llegits, el nombre de variables i els noms de les variables. Recordeu que abans de carregar l'arxiu, cal inspeccionar quin tipus de format csv es tracta per tal que la seva lectura sigui apropiada.

Càrrega de les dades. El separador de camps és una coma i el símbol decimal és un punt. Depenent de la carpeta:

```
satisfaccioLaboral <-
  read.table("D:/Users/cagarcia/uoc/M2.954 - Estadística avançada/PAC1/rawData.csv",
             header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE,
             blank.lines.skip = TRUE)
```

Hi ha 12 variables i 1920 registres (valor obtingut amb la funció nrow). El resum abans de fer cap tractament de dades és:

```
summary(satisfaccioLaboral)
```

```
##                       city     sex       educ_level      job_type
##  barcelona              :170   f:246   Min.   :1.00   Min.   :1.0
##  Barcelona              :470   F:714   1st Qu.:1.75   1st Qu.:1.0
##  madrid                 :149   m:234   Median :2.50   Median :1.5
##  Madrid                 :491   M:726   Mean   :2.50   Mean   :1.5
##  santiago de compostela:161             3rd Qu.:3.25   3rd Qu.:2.0
##  Santiago de Compostela:479             Max.   :4.00   Max.   :2.0
##
##    happiness         age          seniority       sick_leave
##  5.23   :  23   Min.   :18.00   Min.   : 0.00   Min.   : 0.000
##  6.23   :  23   1st Qu.:33.00   1st Qu.:17.00   1st Qu.: 0.000
##  5.33   :  19   Median :40.00   Median :20.00   Median : 0.000
##  6.33   :  19   Mean   :40.33   Mean   :19.99   Mean   : 6.667
##  5.13   :  18   3rd Qu.:47.00   3rd Qu.:24.00   3rd Qu.: 0.000
##  (Other):1806   Max.   :67.00   Max.   :35.00   Max.   :78.000
##  NA's   :  12
##   sick_leave_b      work_hours      Cho_initial     Cho_final
##  Min.   :0.0000   Min.   :-44.70   Min.   :0.95   Min.   :0.990
##  1st Qu.:0.0000   1st Qu.: 35.10   1st Qu.:1.15   1st Qu.:1.250
##  Median :0.0000   Median : 37.90   Median :1.25   Median :1.350
##  Mean   :0.2464   Mean   : 34.49   Mean   :1.20   Mean   :1.351
##  3rd Qu.:0.0000   3rd Qu.: 40.70   3rd Qu.:1.25   3rd Qu.:1.450
##  Max.   :1.0000   Max.   : 88.00   Max.   :1.45   Max.   :1.680
##
```

El tipus s'ha obtingut amb la funció class (només carregant les dades, sense cap tractament).

```
sapply(satisfaccioLaboral, class)
```

```
##         city          sex   educ_level     job_type    happiness
##     "factor"     "factor"    "integer"    "integer"     "factor"
##          age    seniority   sick_leave sick_leave_b   work_hours
##    "integer"    "integer"    "integer"    "integer"    "numeric"
##  Cho_initial    Cho_final
##    "numeric"    "numeric"
```

Table 1: Tipus de cada variable

Variable	**Tipus**
city	factor
sex	factor
educ_level	integer
job_type	integer
happiness	factor
age	integer
seniority	integer
sick_leave	integer
sick_leave_b	integer
work_hours	numeric
Cho_initial	numeric
Cho_final	numeric

2. Indiqueu el tipus de variable estadística de cada variable.

Table 2: Tipus de variable estadística

Variable	**Tipus**
city	categòrica nominal
sex	categòrica nominal
educ_level	categòrica ordinal
job_type	categòrica ordinal
happiness	quantitativa contínua
age	quantitativa discreta
seniority	quantitativa discreta
sick_leave	quantitativa discreta
sick_leave_b	categòrica ordinal
work_hours	quantitativa contínua
Cho_initial	quantitativa contínua
Cho_final	quantitativa contínua

1. **City.** Ciutat de la persona treballadora. És nominal ja que conté informació alfanumèrica.
2. **Sex.** Sexe de la persona treballadora. És nominal ja que conté informació alfanumèrica.
3. **Educ_level.** Nivell d'educació de la persona treballadora. Conté informació numèrica que amaga informació categòrica (les categories es classifiquen com a 1, 2, 3 o 4).
4. **Job_type.** Tipus de treball de la persona treballadora. Conté informació numèrica que amaga informació categòrica (les categories es classifiquen com a 1 o 2). L'enunciat fa servir les classificacions C: qualificat i PC: poc qualificat.

5. **Happiness.** Nivell de felicitat de la persona treballadora. Conté numeració contínua.
6. **Age.** Edat de la persona treballadora. Aquesta variable podria ser considerada com a quantitativa contínua (si pensem que podem expressar l'edat amb decimals). Jo considero que la naturalesa de la variable és discreta (es respon 41 anys, no 41.3).
7. **Seniority.** Anys d'experiència de la persona treballadora. Mateix cas que la variable Age.
8. **Sick_leave.** Dies en que la persona treballadora ha estat malalta en el darrer any. La considero quantitativa discreta perquè els dies s'indiquen de forma sencera. El dia o es conta o no es conta.
9. **Sick_leave_b.** Variable binària que indica si la persona ha estat malalta. Evidentment, això és una categòrica ordinal.
10. **Work_hours.** Hores setmanals de feina de promig de la persona. És quantitativa contínua.
11. **Cho_initial.** Nivell de colesterol de la persona a la penúltima revisió mèdica. És quantitativa contínua.
12. **Cho_final.** Nivell de colesterol de la persona a la darrera revisió mèdica. És quantitativa contínua.

3. En el cas que R no assigni el tipus apropiat a alguna variable, realitzeu la conversió necessària per tal que el tipus final de cada variable sigui l'apropiat. Per exemple, possibles errors de variables quantitatives amb confusió en el separador decimal.

Table 3: Tipus de variable estadística

Variable	Tipus assignat	Tipus objectiu
city	factor	factor
sex	factor	factor
educ_level	integer	ordered
job_type	integer	factor
happiness	factor	numeric
age	integer	integer
seniority	integer	integer
sick_leave	integer	integer
sick_leave_b	integer	logical
work_hours	numeric	numeric
Cho_initial	numeric	numeric
Cho_final	numeric	numeric

Nivell d'educació

La transformació és molt senzilla. Transformem els valors actuals (1, 2, 3 i 4) en els codis indicats a l'enunciat. No tractem els valors buits perquè no n'hi ha.

```
satisfaccioLaboral$educ_level <- factor(satisfaccioLaboral$educ_level,
                                        levels=c(1,2,3,4), labels = c("N", "P", "S", "U"))
table(satisfaccioLaboral$educ_level)
```

```
##
##   N   P   S   U
## 480 480 480 480
```

Tipus de feina

La transformació és molt senzilla. Transformem els valors actuals (1 i 2) en codis indicats a l'enunciat (C per a Qualificat i PC per a poc qualificat) . No tractem els valors buits perquè no n'hi ha.

```
satisfaccioLaboral$job_type <- factor(satisfaccioLaboral$job_type, levels=c(1,2),
                                      labels=c("C", "PC"))
table (satisfaccioLaboral$job_type)
```

```
##
##   C  PC
## 960 960
```

Nivell de felicitat

Les dades tenen números amb "," en lloc de "." com a separador de decimals. Substituïm els valors i convertim el tipus. Per ara, mantindrem els valors NA

```
satisfaccioLaboral$happiness <- gsub(",", ".", satisfaccioLaboral$happiness);
satisfaccioLaboral$happiness <- as.numeric(satisfaccioLaboral$happiness);
summary(satisfaccioLaboral$happiness);
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.930   4.930   5.930   5.891   6.830  10.000      12
```

Binari de malaltia

Ara mateix el valor de la columna és 0 (false), 1 (true); convertim a booleà.

```
satisfaccioLaboral$sick_leave_b <- factor(satisfaccioLaboral$sick_leave_b,
                                          levels=c(0,1), labels=c(FALSE, TRUE));
summary(satisfaccioLaboral$sick_leave_b);
```

```
## FALSE  TRUE
##  1447   473
```

Hores de feina

Podem veure que hi ha valors negatius en les hores de feina. Evidentment, això no és possible. No es poden fer un número negatiu de hores. Els valors semblen molt raonables, amb el signe equivocat.

```
satisfaccioLaboral$work_hours[satisfaccioLaboral$work_hours <= 0]
```

```
##  [1] -31.00004 -37.60014 -39.60019 -41.30002 -41.90006 -39.70012 -36.40013
##  [8] -38.10003 -35.70020 -37.40001 -34.20018 -41.50005 -37.10011 -38.00019
## [15] -36.10003 -39.50010 -41.00014 -37.10006 -36.00019 -39.30019 -36.70007
## [22] -42.00016 -35.50011 -36.80002 -42.20018 -41.00007 -42.60010 -35.00017
## [29] -38.10004 -41.20006 -40.80003 -39.20011 -41.30000 -38.60014 -42.80008
## [36] -42.30006 -33.40013 -41.10010 -43.60007 -33.90006 -34.10013 -34.70000
## [43] -39.20006 -39.00018 -34.20016 -39.30011 -35.30020 -40.80019 -39.80015
## [50] -34.40015 -38.70017 -40.70004 -39.50019 -38.40014 -42.90009 -34.80007
## [57] -38.50000 -36.50009 -30.00012 -30.80002 -37.50003 -38.10009 -35.50011
## [64] -26.20016 -40.30012 -39.50014 -37.30002 -43.80012 -42.40018 -37.70009
## [71] -40.30012 -36.60017 -34.70001 -37.00001 -43.00007 -39.70018 -39.20020
## [78] -35.20009 -33.30013 -43.10018 -37.30006 -28.40006 -31.30015 -37.10008
## [85] -38.80008 -39.10001 -38.40004 -38.30013 -35.50009 -37.60005 -37.30012
## [92] -39.50007 -36.40001 -42.10007 -44.70013 -36.30003
```

Podem pensar que es un error d'introducció de dades i que el signe és erroni. Canviem el signe multiplicant per -1.

```
satisfaccioLaboral$work_hours <- -1 *
  satisfaccioLaboral$work_hours[satisfaccioLaboral$work_hours <= 0]
```

Resum de les variables

Convertim les variables que encara no estan en el tipus correcte.

```
satisfaccioLaboral$sick_leave_b <- as.logical(satisfaccioLaboral$sick_leave_b)
satisfaccioLaboral$educ_level <- as.ordered(satisfaccioLaboral$educ_level)
```

Podem veure que ja tenim els tipus de dades objectiu

```
sapply(satisfaccioLaboral, class)
```

```
## $city
## [1] "factor"
##
## $sex
## [1] "factor"
##
## $educ_level
## [1] "ordered" "factor"
##
## $job_type
## [1] "factor"
##
## $happiness
## [1] "numeric"
##
## $age
## [1] "integer"
##
## $seniority
## [1] "integer"
##
## $sick_leave
## [1] "integer"
##
## $sick_leave_b
## [1] "logical"
##
## $work_hours
## [1] "numeric"
##
## $Cho_initial
## [1] "numeric"
##
## $Cho_final
## [1] "numeric"
```

4. Normalitzeu / Estandaritzeu les variables quantitatives.

Nivell de felicitat

El problema que tenim és que hi ha valors NA a la variable. Tenim 12 casos. Tenim vàries estratègies possibles:

1. Descartar les mostres
2. Assignar un valor en funció d'una variable correlacionada
3. Assignar un valor en funció de la distància de les altres mostres
4. Assignar el valor de la mitja de les mostres

El rati de valors és molt petit (només 12), així que podem assumir que assignar la mitja és correcte i no afectarà gaire a la mostra. Una vegada fet això, normalitzem la variable:

$$z_i = \frac{x_i - \overline{x}}{S}$$

```
satisfaccioLaboral$happiness[is.na(satisfaccioLaboral$happiness)] <-
  mean(satisfaccioLaboral$happiness[!is.na(satisfaccioLaboral$happiness)])
satisfaccioLaboral$happinessNorm <- (satisfaccioLaboral$happiness -
  mean(satisfaccioLaboral$happiness)) / sd(satisfaccioLaboral$happiness)
summary(satisfaccioLaboral$happinessNorm)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.94595 -0.70337  0.02184  0.00000  0.69870  3.05657
```

```
sd(satisfaccioLaboral$happinessNorm)
```

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

Com a resultat tenim una variable amb mitjana 0 i desviació típica 1.

Altres variables

Primer estandarizem les working hours a un decimal

```
satisfaccioLaboral$work_hours <- round(satisfaccioLaboral$work_hours, digits = 1)
```

Nomalitzem les altres variables de la mateixa forma, que no tenen valors NA.

```
satisfaccioLaboral$ageNorm <-
  (satisfaccioLaboral$age - mean(satisfaccioLaboral$age)) /
  sd(satisfaccioLaboral$age);
satisfaccioLaboral$seniorityNorm <-
  (satisfaccioLaboral$seniority -
     mean(satisfaccioLaboral$seniority)) /
  sd(satisfaccioLaboral$seniority);
satisfaccioLaboral$sick_leaveNorm <-
  (satisfaccioLaboral$sick_leave - mean(satisfaccioLaboral$sick_leave)) /
  sd(satisfaccioLaboral$sick_leave);
satisfaccioLaboral$work_hoursNorm <-
  (satisfaccioLaboral$work_hours - mean(satisfaccioLaboral$work_hours)) /
  sd(satisfaccioLaboral$work_hours);
satisfaccioLaboral$Cho_initialNorm <-
  (satisfaccioLaboral$Cho_initial - mean(satisfaccioLaboral$Cho_initial)) /
  sd(satisfaccioLaboral$Cho_initial);
```

```
satisfaccioLaboral$Cho_finalNorm <-
  (satisfaccioLaboral$Cho_final - mean(satisfaccioLaboral$Cho_final)) /
  sd(satisfaccioLaboral$Cho_final);
```

S'han creat columnes noves amb els valors normalitzats per tal de no perdre la informació original. Crec que és millor crear columnes noves per tal de fer l'anàlisi i no deixar de tenir les columnes originals.

5. Normalitzeu / Estandaritzeu les variables qualitatives.

Ciutat

Per a normalitzar, primer llevem els espais en blanc. Després passem totes les paraules a TitleCase (primera lletra en majúscula i les altres en minúscula). Finalment convertim les preposicions a minúscules (en el nostre exemple, només tenim el "de").

```
satisfaccioLaboral$city <- trimws(satisfaccioLaboral$city, "both");
satisfaccioLaboral$city <- tools::toTitleCase(satisfaccioLaboral$city);
satisfaccioLaboral$city <- stringr::str_replace(satisfaccioLaboral$city, 'De', 'de');
table(satisfaccioLaboral$city);
```

```
##
##              Barcelona                 Madrid Santiago de Compostela
##                    640                    640                    640
```

Sexe

Tenim la variable en majúscules i minúscules. La transformem en majúscules, a més de llevar-li els espais en banc.

```
satisfaccioLaboral$sex <- toupper(trimws(satisfaccioLaboral$sex, "both"));
table(satisfaccioLaboral$sex);
```

```
##
##   F   M
## 960 960
```

6. Reviseu possibles inconsistències entre variables: age versus seniority i número d'hores setmanals promig amb valor negatiu.

Les hores promig ja han estat modificades al punt 3 per a poder obtenir valors correctes a la normalització. Recordem que hem multiplicat per -1 els valors negatius.

```
summary(satisfaccioLaboral$work_hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.20   36.08   38.20   37.99   40.40   44.70
```

Per a trobar incoherències entre les variables age i seniority, primer crearem una variable amb l'edat en que la persona va començar a fer feina (age - seniority).

```
satisfaccioLaboral$startAge <- satisfaccioLaboral$age - satisfaccioLaboral$seniority
```

I mirem quants tenien menys de 18 anys quan varen començar a treballar (en principi són dades inconsistents); en tenim 543.

```r
nrow(subset(satisfaccioLaboral, startAge < 18,))
```

```
## [1] 543
```

Assignem els valors a seniority tal que si li restem a l'edat mai sigui inferior a 18.

```r
satisfaccioLaboral$seniority[satisfaccioLaboral$startAge < 18] <-
  satisfaccioLaboral$seniority[satisfaccioLaboral$startAge < 18] -
  (18 - satisfaccioLaboral$startAge[satisfaccioLaboral$startAge < 18])
```

Tornem a calcular la variable startAge com age - seniority i validem que ja no hi ha cap valor inconsistent:

```r
satisfaccioLaboral$startAge <- satisfaccioLaboral$age - satisfaccioLaboral$seniority
summary(satisfaccioLaboral$startAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0    18.0    20.0    21.3    24.0    35.0
```

El resum de seniority és:

```r
summary(satisfaccioLaboral$startAge)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.0    18.0    20.0    21.3    24.0    35.0
```

Recalculem la variable normalitzada de seniority:

```r
satisfaccioLaboral$seniorityNorm <-
  (satisfaccioLaboral$seniority - mean(satisfaccioLaboral$seniority)) /
  sd(satisfaccioLaboral$seniority)
summary(satisfaccioLaboral$seniorityNorm)
```

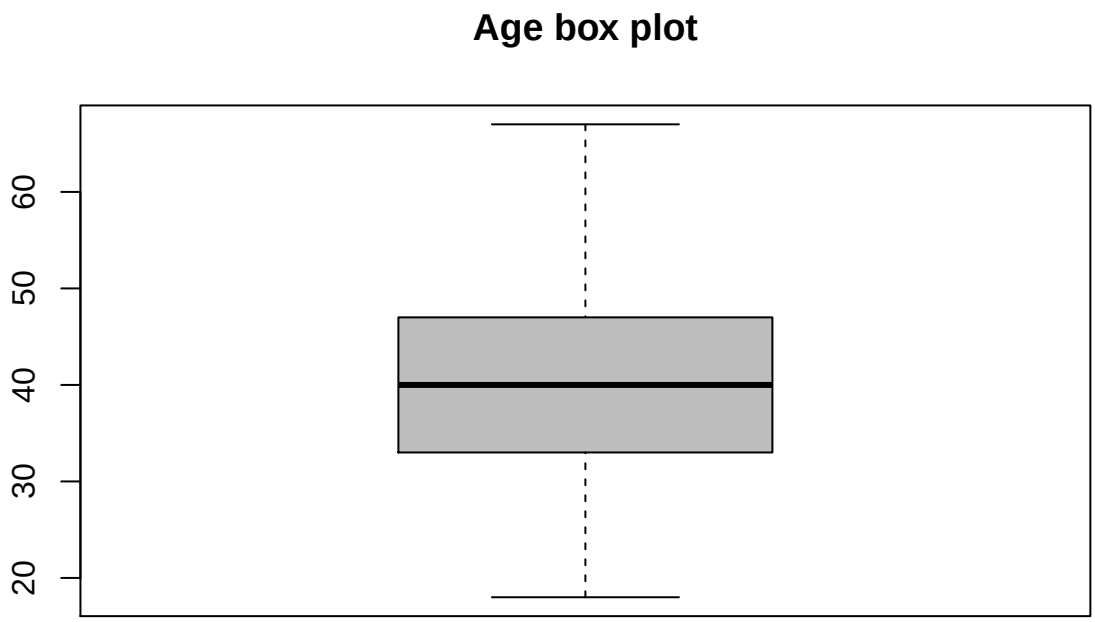
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.7788 -0.5882  0.1420  0.0000  0.7262  2.3327
```

7. Busqueu valors atípics en les variables quantitatives.

i. Presenteu un boxplot per cada variable quantitativa.

Edat

```r
boxplot(satisfaccioLaboral$age,main="Age box plot", col="gray")
```



No hi ha valors outliers:

```
boxplot.stats(satisfaccioLaboral$age)$out
```

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

Ens assegurem calculant el rang d'edat en base al rang interquantil:

```
quantile(satisfaccioLaboral$age, 0.25) -
  (IQR(satisfaccioLaboral$age) * 1.5)
```

```
## 25%
##  12
```

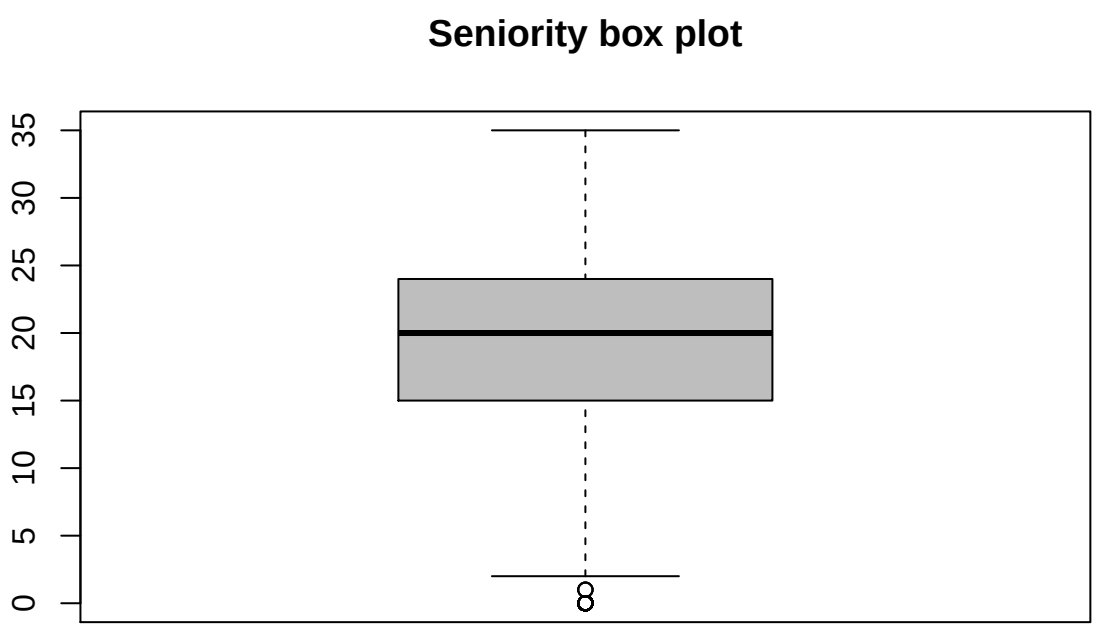
```
quantile(satisfaccioLaboral$age, 0.75) +
  (IQR(satisfaccioLaboral$age) * 1.5)
```

```
## 75%
##  68
```

Efectivament, no hi ha ningú amb menys de 12 anys ni amb més de 68.

Anys d'experiència

```
boxplot(satisfaccioLaboral$seniority,main="Seniority box plot", col="gray")
```



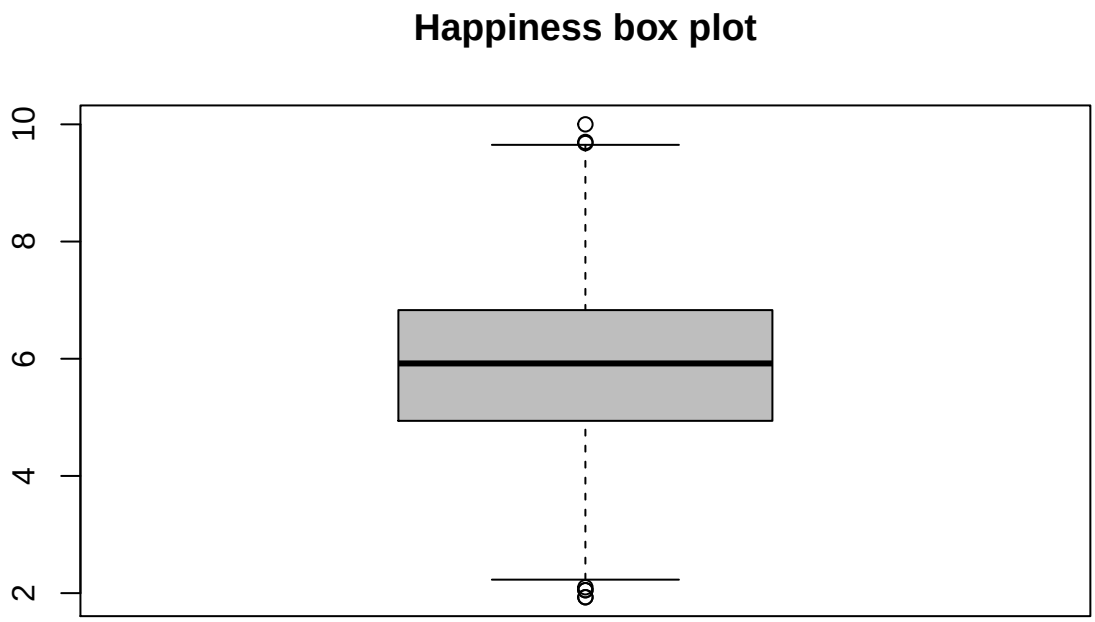

```
boxplot.stats(satisfaccioLaboral$seniority)$out
```

```
##  [1] 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
## [36] 0 0 0 0 0 0 0 1 0 0 0 0
```

Aquí sí hi ha valors outliers, els que tenen 0 o 1 any d'experiència

Nivell de felicitat

```
boxplot(satisfaccioLaboral$happiness,main="Happiness box plot", col="gray")
```




```
boxplot.stats(satisfaccioLaboral$happiness)$out
```

```
## [1] 10.00  2.10  9.70  9.68  2.05  2.05  1.93  1.93
```

Aquí tenim valors outliers tant per abaix com per adalt.

Dies de baixa

Aquí tenim molt de valors outliers; la gran majoria de valors són 0. Això provoca que tant el primer quantil com el tercer siguin 0. Això fa que qualsevol valor diferent de 0 sigui outlier.

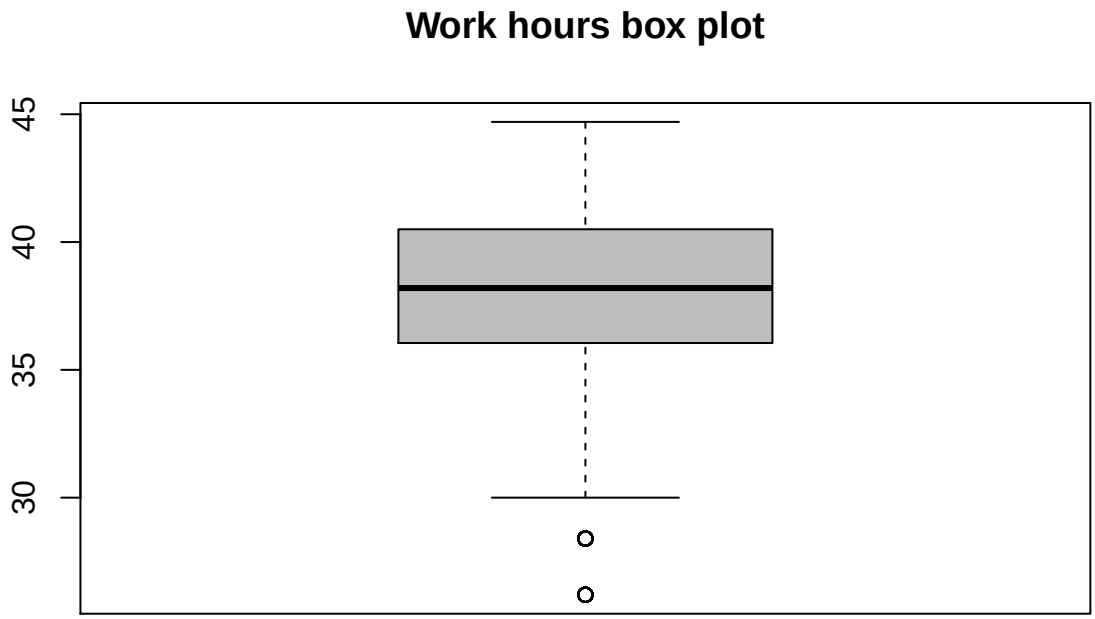
```
boxplot(satisfaccioLaboral$sick_leave,main="Sick leave box plot", col="gray")
```


```
boxplot.stats(satisfaccioLaboral$sick_leave)$out
```

```
##   [1]  6 24 36  1 23  9 53 42 14 27 50 26  3 18 16 10 48 32 51 35 40 28 32
##  [24] 38 51 20 32 41 41 13 38 39 23 36 13 31 30 32 46  8 10 38 56 19 46 33
##  [47] 35  6  9 24  9 11 52 20 15 30 22 22 43  4 34 29 67 40 21  4 26 49  9
##  [70] 16 29 37 27 39 21 31 53 11 22 26 36 14 41 43 12 15  3 12 22 25 53 26
##  [93] 15 36 20 22 29 69 40 34 22 33 14 55 57 20  6 16 40 34 15 47  8 17 40
## [116] 60 68 27 21 45  5 57 34 18 33 11  8  6 20 26 39 20 25 27  1 51 31  3
## [139] 14  7 13 59  7 11 16 29  8 24 57 48 53 27 20 26 27 12 29 33 39 32 45
## [162] 21 56  1 20 41 18 33 23 20 18 41 51 18 57 40 27 12 15  6 22 16 28 17
## [185] 23 53 12  3 39  7 32 30 38 17 14 72 26 28 15 45  6 37 49 22  9 16 21
## [208] 39 19  1 19 31  5 40 20 18 11 44  8 14 46 16 24 42 30 46  1 17 19 24
## [231] 15 25 35 45  7  9 43  8 51 64 21 26 21 48 29 12 19 18 49 42 30 33 24
## [254]  8 14 14  7 24 31 33 17 54 56 32  4  5 34 11 34 27 56 38 11 38 29  7
## [277] 22 30 15  2  4 21 39 19 20 51 31 22 15 17 11 11 28 40 11 35 56 15 44
## [300] 29 29 38 56 29 42 34 11 15 31 21 17 43 23 39  6 40 57  7 33 23  3 24
## [323] 35  6  5 11  6  5 45 40 52 25 25 20 38 66 24 42 39 17 40 19 27 18 45
## [346] 23  3 34 46 15 23 27 52 22 22  8 39 16 17 12 20 16 13 23 25 41  4 38
## [369] 18  7 52 21 38 49 35 22 78 29 12  5 37 22 12  6 42 28 34 11 21 14 14
## [392] 32 56 10 60 27 11 30 41 22  3 45  2 35 16 45 47 34 27 18 31 46 42 11
## [415] 40  4 29 28 13 17 29 50  6 47 23  3 11 22 36 35 30 16 29 28 42 32  8
## [438] 21 58 19 31 29 23 37 34 16 10 32 26 47 15 35 22 14 14 17 27  9 23 26
## [461] 29 33 57 18 20 31 39 21 37 25 45 18 47
```

Hores de feina

```
boxplot(satisfaccioLaboral$work_hours,main="Work hours box plot", col="gray")
```




```
boxplot.stats(satisfaccioLaboral$work_hours)$out
```

```
##  [1] 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4
## [15] 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4
## [29] 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4 26.2 28.4
```

Tenim 40 valors fora dels límits (tots per abaix)

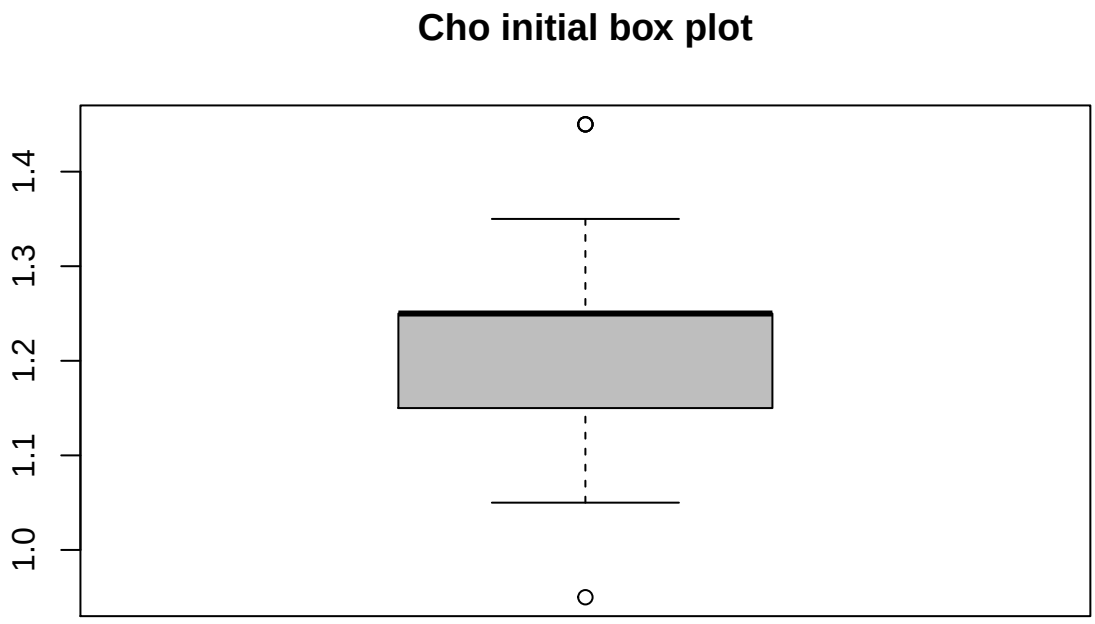
```
knitr::kable(head(satisfaccioLaboral), format = "latex")
```

city	sex	educ_level	job_type	happiness	age	seniority	sick_leave	sick_leave_b	work_hours	Cho_
Barcelona	M	N	C	7.230000	45	23	0	FALSE	31.0	
Barcelona	M	N	C	5.930000	36	18	0	FALSE	37.6	
Barcelona	M	N	C	5.890639	34	16	0	FALSE	39.6	
Barcelona	M	N	C	6.330000	37	18	0	FALSE	41.3	
Barcelona	M	N	C	8.930000	35	17	0	FALSE	41.9	
Barcelona	M	N	C	7.230000	29	11	6	TRUE	39.7	

Els resultats són els mateixos si treballem amb la variable normalitzada

Colesterol penúltima medició

```
boxplot(satisfaccioLaboral$Cho_initial,main="Cho initial box plot", col="gray")
```




```
boxplot.stats(satisfaccioLaboral$Cho_initial)$out
```

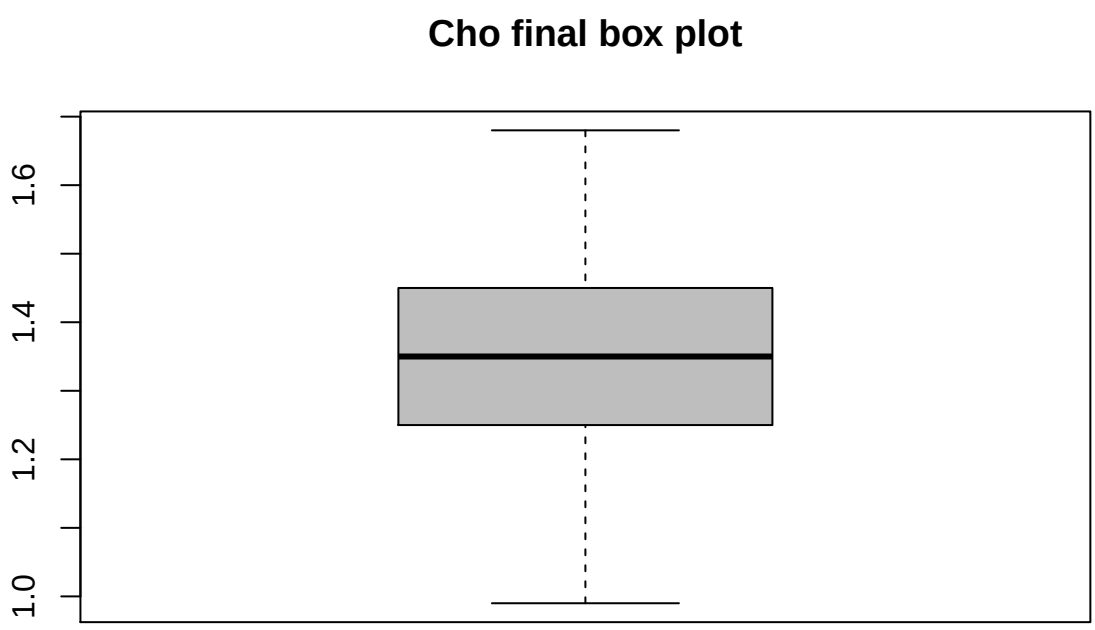
```
## [1] 1.45 1.45 0.95 1.45
```

Aquí també tenim valors outliers tant per adalt com per abaix

Colesterol última medició

Aquí no tenim valors outliers, perquè el rang interquantílic és més ampli. És a dir: els valors són més lluny de la mitja. Això fa que valors que amb la gràfica anterior s'haguessin considerat outliers, aquí són "normals".

```
boxplot(satisfaccioLaboral$Cho_final,main="Cho final box plot", col="gray")
```




```
boxplot.stats(satisfaccioLaboral$Cho_final)$out
```

```
## numeric(0)
```

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

variables	happiness	age	seniority	sick_leave	work_hours	cho_ini	cho_fin
Mitjana aritmètica	5.890639	40.331250	19.027604	6.666667	37.986458	1.2002083	1.3513229
Mediana	5.920000	40.000000	20.000000	0.000000	38.200000	1.2500000	1.3500000
Mitjana retallada (al 5%)	5.892238	40.247107	19.308449	4.640046	38.158565	1.2001157	1.3512789
Mitjana winsoritzada (al 5%)	5.888439	40.274896	19.127604	6.176042	38.042708	1.2001042	1.3516510
Rang Interquartílic	1.885000	14.000000	9.000000	0.000000	4.325000	0.1000000	0.2000000
Desviació estàndard	1.344434	9.879243	6.847309	13.899975	3.443904	0.0771425	0.1279922
DAM	1.363992	10.378200	5.930400	0.000000	3.187590	0.1482600	0.1482600

Els resultats són els mateixos si treballem amb la variable normalitzada

8. Valors perduts

i.Busqueu quines variables i quins registres tenen valors perduts.

L'única variable amb valors perduts és happiness (12 NA). Com abans he assignat el valor de la mitja per a fer l'anàlisi, torn enrere el canvi:

```
satisfaccioLaboralAux <-
  read.table("D:/Users/cagarcia/uoc/M2.954 - Estadística avançada/PAC1/rawData.csv",
```

```
            header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE,
            blank.lines.skip = TRUE)
satisfaccioLaboral$happiness <- satisfaccioLaboralAux$happiness
satisfaccioLaboral$happiness <- gsub(",", ".", satisfaccioLaboral$happiness);
satisfaccioLaboral$happiness <- as.numeric(satisfaccioLaboral$happiness);
head(satisfaccioLaboral$happiness)
```

```
## [1] 7.23 5.93   NA 6.33 8.93 7.23
```

Els registres NA són:

	city	sex	age	seniority
3	Barcelona	M	34	16
7	Barcelona	M	26	8
12	Barcelona	M	31	13
55	Barcelona	M	48	25
66	Barcelona	M	30	12
76	Barcelona	M	35	16
100	Barcelona	M	43	23
200	Barcelona	M	50	23
300	Barcelona	M	31	13
800	Madrid	M	33	15
1000	Madrid	F	41	19
1050	Madrid	F	52	27

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

Prendrem com a nombre K = 5, que és el valor per defecte. Considerem totes les variables, no s'exclou cap columna. Per veure el resultat, guardem els valors a la variable happinessComplete. Així podrem comprovar els valors assignats:

```
satisfaccioLaboral.complet <- VIM::kNN(satisfaccioLaboral);
```

```
## Warning in gowerD(don_dist_var, imp_dist_var, weights = weightsx,
## numericalX, : NAs introduced by coercion

## Warning in gowerD(don_dist_var, imp_dist_var, weights = weightsx,
## numericalX, : NAs introduced by coercion

## Warning in gowerD(don_dist_var, imp_dist_var, weights = weightsx,
## numericalX, : NAs introduced by coercion

## Warning in gowerD(don_dist_var, imp_dist_var, weights = weightsx,
## numericalX, : NAs introduced by coercion
```

```
satisfaccioLaboral$happinessComplete <- satisfaccioLaboral.complet$happiness;
```

Comprovem els valors assignats:

```
satisfaccioLaboral$happinessComplete[is.na(satisfaccioLaboral$happiness)]
```

```
##  [1] 5.68 5.35 6.61 5.03 5.38 6.23 6.93 6.03 6.10 6.23 6.33 4.75
```

Ara que sabem que està bé, li assignem a la variable:

```
satisfaccioLaboral$happiness <- satisfaccioLaboral$happinessComplete
```

Si tornem a calcular els estimadors de la variable:

variables	happiness_old	happiness_new
Mitjana aritmètica	5.890639	5.890620
Mediana	5.920000	5.930000
Mitjana retallada (al 5%)	5.892238	5.892216
Mitjana winsoritzada (al 5%)	5.888439	5.888420
Rang Interquartílic	1.885000	1.900000
Desviació estàndard	1.344434	1.345349
DAM	1.363992	1.363992

9. Realitzeu un breu estudi descriptiu de les dades un cop depurades.

Mostrem el summary de totes les variables una vegada depurades. De les que no es veuen les dades, mostrem la distribució mitjançant l'operació table.

```
sapply(satisfaccioLaboral[c("happiness", "educ_level",
                            "job_type", "age", "seniority", "sick_leave",
                            "sick_leave_b", "work_hours", "Cho_initial",
                            "Cho_final")], summary)
```

```
## $happiness
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.930   4.930   5.930   5.891   6.830  10.000
##
## $educ_level
##   N   P   S   U
## 480 480 480 480
##
## $job_type
##   C  PC
## 960 960
##
## $age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   33.00   40.00   40.33   47.00   67.00
##
## $seniority
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   15.00   20.00   19.03   24.00   35.00
##
## $sick_leave
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   6.667   0.000  78.000
##
## $sick_leave_b
##    Mode   FALSE    TRUE
## logical    1447     473
##
## $work_hours
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   26.20   36.08   38.20   37.99   40.40   44.70
##
## $Cho_initial
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.95    1.15    1.25    1.20    1.25    1.45
##
## $Cho_final
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.990   1.250   1.350   1.351   1.450   1.680
```

```
table(satisfaccioLaboral$city)
```

```
##
##              Barcelona                 Madrid Santiago de Compostela
##                    640                    640                    640
```

```
table(satisfaccioLaboral$sex)
```

```
##
##   F   M
## 960 960
```

10. Finalment, emmagatzemeu les dades en un arxiu de dades corregit.

Desem els resultats al fitxer rawDataNet.csv

```
write.table(satisfaccioLaboral[c("city", "sex", "happiness", "educ_level",
  "job_type", "age", "seniority", "sick_leave",
  "sick_leave_b", "work_hours", "Cho_initial",
  "Cho_final")],
  file = "D:/Users/cagarcia/uoc/M2.954 - Estadística avançada/PAC1/rawDataNet.csv",
  append = FALSE, sep = ";", row.names = FALSE)
```