Màster DS UOC. Tipologia i cicle de vida de les dades. Pràctica II.

Ángel Lavado Cuevas

11 de junio, 2018

Índex

1 Presentació

L'objectiu d'aquesta activitat es basa en posar en pràctica els coneixements adquirits durant el màster a l'assignatura Tipologia i cicle de vida de les dades. En concret es realitzarà el tractament d'un dataset seleccionat que permeti aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi presentades a l'assignatura.

Els objectius concrets plantejats per aquesta pràctica són els següents:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar habilitats que permetin continuar adquirint coneixement mitjançant l'auto-aprenentatge.
- Fomentar i desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Amb aquesta pràctica es preten desenvolupar les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adeqüat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi.

2 Descripció del dataset

El dataset que s'utilitzarà durant tota l'activitat serà l'obtingut a la pràctica 1 d'aquesta mateixa assignatura. El conjunt de dades va ser creat mitjançant un procés dissenyat de web-scraping i es troba disponible, així com la resta de documentació relacionada, al següent *link*. Es tracta d'una mostra dels pacients en espera al servei d'urgències del Consorci Sanitari de l'Alt Penedès. Aquesta informació és actualitzada a la web institucional del consorci cada cinc minuts, i aporta el número de pacients que estan sent atesos en aquell moment al seu servei d'urgències, distribuïts pel corresponent nivell de triatge. El nivell de triatge és un mètode de classificació de la gravetat de les urgències, essent els valors 1 i 2 els més urgents i que són acompanyats directament al box d'atenció. Els pacients de nivell 3, després del triatge, entren a una sala d'espera interna on són vigilats constantment pel personal sanitari. Els nivells 4 i 5 corresponen als menys urgents, els quals, un cop avaluats i assignats el nivell passen a la sala d'espera fins que puguin ser cridats al box per a la seva atenció. El dataset inclou una fila corresponent a cada procés de captura finalitzat amb èxit. Es va realitza un procés de scraping programat cada cinc minuts. El període de temps aportat al dataset d'exemple correspon a les extraccions realitzades cada cinc minuts des de les 21:16 hores del

dia 11/04/2018 fins les 23:56 hores del dia 12/04/2018. Els camps que inclou el conjunt de dades són els següents:

dataActualitzacioWeb: data i hora d'actualització de les dades informada per la pròpia web, en format "Dimecres, 11 d'abril de 2018 a les 21'16h".

dataCaptura: data en format yyyy-mm-dd de la data en la qual ha estat capturada la informació de la fila mitjançant el procés de captura.

horaCaptura: hora en format hh:mm de la data en la qual ha estat capturada la informació de la fila mitjançant el procés de captura.

PacientsNivell1: número de pacients en espera en el moment de la captura per al nivell 1 de triatge.

PacientsNivell2: número de pacients en espera en el moment de la captura per al nivell 2 de triatge.

PacientsNivell3: número de pacients en espera en el moment de la captura per al nivell 3 de triatge.

PacientsNivell4: número de pacients en espera en el moment de la captura per al nivell 4 i 5 de triatge.

TempsEsperaAdults: temps d'espera informat del darrer pacient visitat dels nivells 4 i 5 per a adults.

TempsEsperaPediatric: temps d'espera informat del darrer pacient visitat dels nivells 4 i 5 per a pediatria.

Els temps d'espera, tant d'adults com pediàtric, es corresponen als minuts que ha esperat el darrer pacient atès.

El dataset plantejat ens permetrà analitzar tant el número de pacients en espera com els temps que han experimentat aquests pacients durant una sèrie temporal de 24 hores, fraccionades en porcions de 5 minuts. El pacient que és classificat com a nivell 4 o 5 i és indicat a esperar en la sala d'espera per ser atés, si la sala d'espera és molt plena és molt habitual que interpreti que el temps que haurà d'espera serà molt llarg, i al contrari, si la sala d'espera és buida preveu que serà cridat a entrar en un curt espai de temps. Aquesta deducció no contempla que els pacients que estan en la sala d'espera son els menys greus, i que per tant no es coneix el número i la gravetat dels pacients que están sent atesos en aquell moment. Probablement siguin aquests últims els que realment condicionin el temps d'espera dels que estan ubicats a la sala. A priori hem d'entendre que els recursos establerts per atendre les urgències estan ajustats a la demanda de servei per torn. Intentarem respondre a la següent pregunta a partir de les dades que disposem: **Existeix alguna relació entre el temps d'espera dels pacients dels nivells 4 i 5 i el torn de guàrdia (matí, tarda o nit) en que son atesos?**.

No obstant ens trobem amb una problemàtica identificada en les variables disponibles al dataset, i és que els temps d'espera els disposem diferenciats per a pacients adults i pediàtrics, però el número de pacients en espera per als nivells 4 i 5 no els disposem diferenciats en adults i pediàtrics. A partir de les dades d'activitat anual a urgències disponibles en la informació coorporativa de l'entitat sanitària [1], es pot comprovar com durant els últims dos anys el percentatge de visites

[1] http://www.csap.cat/memoria/2016/activitat_2016/urgencies.html

pediàtriques sobre el total de visites a urgències ha estat d'un 15% de mitjana. En base a aquesta dada ponderarem el valor aportat a la variable PacientsNivell4 en un 85%, i arrodonit a l'alça, per estimar el número de pacients de nivell 4 i 5 adults. Aquest nou valor el registrarem en la nova variable PacientsNivell4Adults. La diferència entre el número de pacients a la variable PacientsNivell4 i la variable PacientsNivell4Adults els considerarem pacients pediàtrics estimats i els registrarem en una nova variable que anomenarem PacientsNivell4Pediatrics. D'altra banda les dades que disposem només corresponen a 24 hores d'atenció al servei d'urgències, per la qual cosa les conclusión que extreurem de l'anàlisi i la resposta que donarem a la pregunta plantejada estaran particularitzades a un sol dia d'atenció. Per extreure conclusions més robustes caldria disposar de dades sobre un període de web-scraping molt més ampli. Igualment l'exercici que realitzarem ha de ser vàlid per replicar en datasets de major vòlum de dades i així obtenir conclusions més significatives.

```
# Carreguem el dataset en una variable mitjançant la funció read.csv
bd <- read.csv("CSAP.csv",header=TRUE)
# Presentem una mostra amb els valors de les 2 primeres files del dataset
kable(t(bd[1:2,]))
```

	1	2
dataActualitzacioWeb	Dimecres, 11 d'abril de 2018 a les 21'16h	Dimecres, 11 d'abril de 2018 a les 21'21h
dataCaptura	2018-04-11	2018-04-11
horaCaptura	21:16	21:21
PacientsNivell1	0	0
PacientsNivell2	4	4
PacientsNivell3	8	8
PacientsNivell4	11	12
tempsEsperaAdults	1h. 24min.	1h. 24min.
tempsEsperaPediatrics	1h. 0min.	1h. 0min.

```
# Mostrem un resum estadístic i del tipus de dades del dataset
summary(bd)
```

```
##                              dataActualitzacioWeb     dataCaptura
##  Dijous, 12 d'abril de 2018 a les 12'06h:  2       2018-04-11: 33
##  Dijous, 12 d'abril de 2018 a les 00'01h:  1       2018-04-12:287
##  Dijous, 12 d'abril de 2018 a les 00'06h:  1
##  Dijous, 12 d'abril de 2018 a les 00'11h:  1
##  Dijous, 12 d'abril de 2018 a les 00'16h:  1
##  Dijous, 12 d'abril de 2018 a les 00'21h:  1
##  (Other)                                 :313
##   horaCaptura  PacientsNivell1 PacientsNivell2 PacientsNivell3
##  22:51  :  2   Min.   :0       Min.   :2.000   Min.   : 5.00
##  22:56  :  2   1st Qu.:0       1st Qu.:2.000   1st Qu.: 6.75
##  23:1   :  2   Median :0       Median :3.000   Median :12.50
##  23:11  :  2   Mean   :0       Mean   :3.266   Mean   :13.44
##  23:16  :  2   3rd Qu.:0       3rd Qu.:5.000   3rd Qu.:21.00
```

```
##  23:21   :  2   Max.   :0       Max.   :6.000   Max.   :26.00
##  (Other):308
##  PacientsNivell4    tempsEsperaAdults tempsEsperaPediatrics
##  Min.   : 0.000    0h. 30min.: 47      1h. 0min. :126
##  1st Qu.: 2.000    0h. 31min.: 39      0h. 8min. : 42
##  Median : 7.000    0h. 3min. : 27      0h. 7min. : 36
##  Mean   : 6.359    0h. 12min.: 15      0h. 4min. : 32
##  3rd Qu.:10.000    0h. 5min. : 14      1h. 16min.: 17
##  Max.   :15.000    0h. 13min.: 12      0h. 41min.: 15
##                   (Other)    :166     (Other)    : 52
```

```
variables<-as.matrix(sapply(bd, class))
kable(variables, caption = "Tipus de dades")
```

Taula 2: Tipus de dades

dataActualitzacioWeb	factor
dataCaptura	factor
horaCaptura	factor
PacientsNivell1	integer
PacientsNivell2	integer
PacientsNivell3	integer
PacientsNivell4	integer
tempsEsperaAdults	factor
tempsEsperaPediatrics	factor

3 Integració i selecció de les dades d'interès a analitzar.

Generem la nova variable PacientsNivell4Adults com detallavem a la descripció del dataset, i crearem una variable nova més amb el valor de l'estimació de pacients pediàtrics d'aquests nivells.

```
bd$PacientsNivell4Adults<-as.integer(round((bd$PacientsNivell4*0.85),0))
bd$PacientsNivell4Pediatrics<-as.integer(bd$PacientsNivell4-bd$PacientsNivell4Adults)
```

De totes les variables disponibles, analitzem les variables dataActualitzacioWeb, dataCaptura i horaCaptura. La informació sobre els pacients en espera i el temps d'espera que s'informa a la *web institucional* a la qual pertanyen les dades, s'actualitza, segons s'indica a la mateixa web, cada cinc minuts. El procés de web-scraping es va realitzar cada cinc minuts també. Podem comprovar a l'analitzar aquestes tres variables com, en determinades captures de dades, la hora i minuts del moment de la captura coincideix amb la hora i minuts que la web informa que les dades han estat actualitzades. En d'altres no. Si analitzem la variable horaCaptura podem observar també com el volcat dels valors mitjançant la funció *read.csv* no ha respectat el valor "0" en determinades posicions, convertint-lo en un valor incorrecte. Per exemple a la fila número 10 podem comprovar com el valor emmagatzemat com a horaCaptura és incomplert:

```
kable(bd[10,1:3])
```

	dataActualitzacioWeb	dataCaptura	horaCaptura
10	Dimecres, 11 d'abril de 2018 a les 22'01h	2018-04-11	22:1

Donat l'exposat es generaran les variables dataActualitzacio i horaActualitzacio a partir de les tres variables indicades existents i es prescindirà finalment de les variables dataActualitzacioWeb, dataCaptura i horaCaptura. Comprovarem que els dies informats a dataActualitzacioWeb i dataCaptura siguin els mateixos per a cada fila mitjançant l'ús de les següents funcions amb expresions regulars:

```
str_sub(bd$dataCaptura[1],-2,-1)
```

```
## [1] "11"
```

```
str_sub(substring(bd$dataActualitzacioWeb[1],regexpr(",",bd$dataActualitzacioWeb[1])),3,
3-nchar(substring(bd$dataActualitzacioWeb[1],regexpr(",",bd$dataActualitzacioWeb[1]))))
```

```
## [1] "11"
```

```
#Amb aquest bucle comprovem si a alguna fila les dues dates son diferents en dia:
for ( i in 1:nrow(bd))
{if(str_sub(bd[i,2],-2,-1)!=str_sub(substring(bd[i,1],regexpr(",",bd[i,1])),
    3,3-nchar(substring(bd[i,1],regexpr(",",bd[i,1])))))
{print("data captura diferent a data actualització")}
}
```

No es localitza cap fila, per tant a totes les files el dia informat com a data de captura és el mateix que el dia informat com a data d'actualització, i podem utilitzar el valor aportat per la variable dataCaptura per informar la nova variable dataActualització. Així doncs generarem una nova variable horaActualitzacio a partir de la informació continguda a la variable dataActualitzacioWeb.

```
#Canviem el nom de la variable dataCaptura
names(bd)<-gsub("dataCaptura","dataActualitzacio",names(bd))
#Creem la variable horaActualitzacio i l'omplim amb la informació de l'hora i minuts que
#disposem a la variable dataActualitzacioWeb
bd$horaActualitzacio<-substr(gsub("'",":",sub(".*\\s+", "", bd[,1])),1,5)
#Mostrem un parell d'exemples de com queden les variables
kable(t(bd[10:11,]))
```

	10	11
dataActualitzacioWeb	Dimecres, 11 d'abril de 2018 a les 22'01h	Dimecres, 11 d'abril de 2018 a les 22'06h
dataActualitzacio	2018-04-11	2018-04-11
horaCaptura	22:1	22:6
PacientsNivell1	0	0
PacientsNivell2	4	4
PacientsNivell3	9	9
PacientsNivell4	9	10
tempsEsperaAdults	0h. 39min.	1h. 44min.
tempsEsperaPediatrics	1h. 16min.	1h. 16min.
PacientsNivell4Adults	8	8
PacientsNivell4Pediatrics	1	2
horaActualitzacio	22:01	22:06

Un cop disposem del valor de la variable horaActualitzacio normalitzat i correctament informat podem crear la variable Torn en base a l'hora que ens informa aquesta varible. Establirem tres torns diferenciats en base als estàndars d'horaris laborals dels col · lectius implicats en aquest sector. En concret definirem un horari de *Mati* quan l'hora d'actualització estigui compresa entre les 07:00 i les 14:59 h., un torn de *Tarda* quan l'hora estigui entre les 15:00 h. i les 22:59 i *Nit* per la resta.

```
for ( i in 1:nrow(bd)){bd$Torn[i]<-if (as.integer(substr(bd[i,12],1,2))>=7
& as.integer(substr(bd[i,12],1,2))<15){"Mati"} else if (substr(bd[i,12],1,2)>=15
& substr(bd[i,12],1,2)<23){"Tarda"}else {"Nit"}}
```

Descartarem per a continuar el nostre anàlisi les variables dataActualitzacioWeb i horaCaptura. La resta de variables del dataset seràn seleccionades i analitzades.

4 Neteja de les dades.

Per treballar més còmodament eliminarem del dataset les variables que ja hem identificat amb les que no treballarem finalment. Comprovem el tipus de dades que ens han quedat després del canvis realitzats.

```
bd<-bd[,-c(1,3)]
variables<-as.matrix(sapply(bd, class))
kable(variables, caption = "Tipus de dades")
```

Taula 5: Tipus de dades

dataActualitzacio	factor
PacientsNivell1	integer
PacientsNivell2	integer
PacientsNivell3	integer
PacientsNivell4	integer
tempsEsperaAdults	factor
tempsEsperaPediatrics	factor
PacientsNivell4Adults	integer
PacientsNivell4Pediatrics	integer
horaActualitzacio	character
Torn	character

Podem comprovar com la incorporació d'alguns valors mitjançant la funció *read.csv* ha convertit algunes variables en un tipus de variable no esperat.

- La variable dataCaptura renombrada com a dataActualització ha estat identificada com a tipus factor per R.
- Les modificacions realitzades sobre la variable dataActualitzacioWeb per generar la variable horaActualitzacio han facilitat que R la reconogui com a tipus character.
- La variable dataActualitzacio cal convertir-la en tipus Date per tal de facilitar la seva anàlisi posteriorment.

```
bd$dataActualitzacio<-as.Date(bd$dataActualitzacio)
class(bd$dataActualitzacio)
```

```
## [1] "Date"
```

La resta de variables han estat identificades per R en el domini esperat. No obstant, farem una conversió de les variables tempsEsperaAdults i tempsEsperaPediatrics a integer, convertint el valor que aporten en format h. i min. a minuts totals i informant del valor total de minuts d'espera en dos noves variables continues minutstempsEsperaAdults i minutstempsEsperaPediatrics.

```
bd$minutstempsEsperaAdults<-as.integer(as.integer(substr(gsub("h. ","",
bd$tempsEsperaAdults),0,2))*60+as.integer(ifelse((substr(substr(
gsub("h. ","",bd$tempsEsperaAdults),3,4),2,2))=="m",substr(substr(gsub("h. ","",
bd$tempsEsperaAdults),3,4),1,1),substr(substr(gsub("h. ","",
bd$tempsEsperaAdults),3,4),1,2))))
```

```
bd$minutstempsEsperaPediatrics<-as.integer(as.integer(substr(gsub("h. ","",
bd$tempsEsperaPediatrics),0,2))*60+as.integer(ifelse((substr(substr(gsub("h. ","",
bd$tempsEsperaPediatrics),3,4),2,2))=="m",substr(substr(gsub("h. ","",
bd$tempsEsperaPediatrics),3,4),1,1),substr(substr(gsub("h. ","",
bd$tempsEsperaPediatrics),3,4),1,2))))
```

Generarem una nova variable més que contingui la data i l'hora d'actualització en format POSIXct per poder graficar temporalment d'una forma més còmode.

```
bd$dataActualitzacioCompleta <- as.POSIXct(paste(as.factor(bd$dataActualitzacio),
                                                 bd$horaActualitzacio, sep=" "))
```

Finalment continuarem l'anàlisi amb les següent variables:

```
bd<-bd[,-c(6,7)]
variables<-as.matrix(sapply(bd, class))
kable(variables, caption = "Tipus de dades")
```

Taula 6: Tipus de dades

dataActualitzacio	Date
PacientsNivell1	integer
PacientsNivell2	integer
PacientsNivell3	integer
PacientsNivell4	integer
PacientsNivell4Adults	integer
PacientsNivell4Pediatrics	integer
horaActualitzacio	character
Torn	character
minutstempsEsperaAdults	integer
minutstempsEsperaPediatrics	integer
dataActualitzacioCompleta	c("POSIXct", "POSIXt")

4.1 Valors zeros o elements buits.

Comprovarem si les variables contenen zeros o elements buits. Podem identificar-los a partir de la funció summary:

```
summary(bd)
```

```
##  dataActualitzacio    PacientsNivell1 PacientsNivell2 PacientsNivell3
##  Min.   :2018-04-11   Min.   :0       Min.   :2.000   Min.   : 5.00
##  1st Qu.:2018-04-12   1st Qu.:0       1st Qu.:2.000   1st Qu.: 6.75
##  Median :2018-04-12   Median :0       Median :3.000   Median :12.50
##  Mean   :2018-04-11   Mean   :0       Mean   :3.266   Mean   :13.44
##  3rd Qu.:2018-04-12   3rd Qu.:0       3rd Qu.:5.000   3rd Qu.:21.00
##  Max.   :2018-04-12   Max.   :0       Max.   :6.000   Max.   :26.00
##  PacientsNivell4  PacientsNivell4Adults PacientsNivell4Pediatrics
```

```
##  Min.   : 0.000   Min.   : 0.000        Min.   :0.000
##  1st Qu.: 2.000   1st Qu.: 2.000        1st Qu.:0.000
##  Median : 7.000   Median : 6.000        Median :1.000
##  Mean   : 6.359   Mean   : 5.384        Mean   :0.975
##  3rd Qu.:10.000   3rd Qu.: 8.000        3rd Qu.:2.000
##  Max.   :15.000   Max.   :13.000        Max.   :2.000
##  horaActualitzacio      Torn           minutstempsEsperaAdults
##  Length:320         Length:320         Min.   :  1
##  Class :character   Class :character   1st Qu.: 13
##  Mode  :character   Mode  :character   Median : 30
##                                        Mean   : 31
##                                        3rd Qu.: 42
##                                        Max.   :104
##  minutstempsEsperaPediatrics dataActualitzacioCompleta
##  Min.   : 2.00               Min.   :2018-04-11 21:16:00
##  1st Qu.: 8.00               1st Qu.:2018-04-12 03:54:45
##  Median :41.00               Median :2018-04-12 10:33:30
##  Mean   :36.33               Mean   :2018-04-12 10:33:37
##  3rd Qu.:60.00               3rd Qu.:2018-04-12 17:12:15
##  Max.   :76.00               Max.   :2018-04-12 23:56:00
```

El resum estadístic ja ens informa que les variables PacientsNivell1, PacientsNivell4, PacientsNivell4Adults i PacientsNivell4Pediatrics contenen valors 0. No ens informa de l'existència de valors NAs en el dataset.

Una altre forma d'analitzar si alguna variable conté valors buits o zeros és mitjançant el següent codi:

```
print("No existeix cap valor buit?")
```

```
## [1] "No existeix cap valor buit?"
```

```
sapply(bd, function(x){all(!is.na(x))})
```

```
##           dataActualitzacio             PacientsNivell1
##                        TRUE                        TRUE
##             PacientsNivell2             PacientsNivell3
##                        TRUE                        TRUE
##             PacientsNivell4       PacientsNivell4Adults
##                        TRUE                        TRUE
##   PacientsNivell4Pediatrics           horaActualitzacio
##                        TRUE                        TRUE
##                        Torn     minutstempsEsperaAdults
##                        TRUE                        TRUE
## minutstempsEsperaPediatrics   dataActualitzacioCompleta
##                        TRUE                        TRUE
```

```
print("No existeix cap valor 0?")
```

```
## [1] "No existeix cap valor 0?"
```

```
sapply(bd, function(x){all((x)!=0)})
```

```
##            dataActualitzacio             PacientsNivell1
##                         TRUE                       FALSE
##              PacientsNivell2             PacientsNivell3
##                         TRUE                        TRUE
##              PacientsNivell4       PacientsNivell4Adults
##                        FALSE                       FALSE
##   PacientsNivell4Pediatrics           horaActualitzacio
##                        FALSE                        TRUE
##                         Torn     minutstempsEsperaAdults
##                         TRUE                        TRUE
## minutstempsEsperaPediatrics   dataActualitzacioCompleta
##                         TRUE                        TRUE
```

Com ens informava la funció *summary* no exiteixen valors buits i les variables PacientsNivell1, PacientsNivell4, PacientsNivell4Adults i PacientsNivell4Pediatrics contenen valors zero. El valor zero existent a les variables PacientsNivell1, PacientsNivell4, PacientsNivell4Adults i PacientsNivell4Pediatrics és el valor que correspon a aquestes variables en la fila de captura. Informa de que no existia cap pacient en espera de ser visitat, del nivell indicat, en el moment d'actualitzar les dades. Per tant és un valor 0 a tots els efectes i te la seva significació, aportant la informació detallada anteriorment.

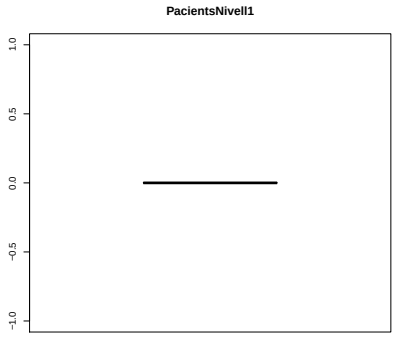
El nostre dataset no conté valors buits. Si haguessim identificat algún valors sense informar en alguna fila, segurament hagués estat donat a algun error en la captura de la fila o tupla. Caldria analitzar en detall com tractar aquestes dades en base a la variable que es localitzessin i la seva rellevància per l'anàlisi global, però a priori l'estratègia a seguir en aquest cas hagués estat eliminar la fila complerta.

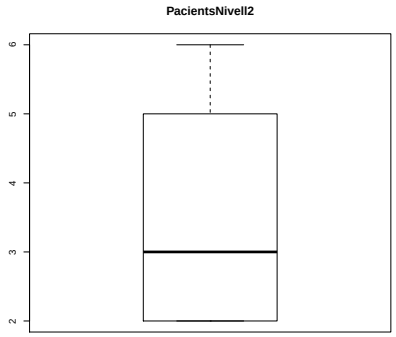
4.2 Identificació i tractament de valors extrems.

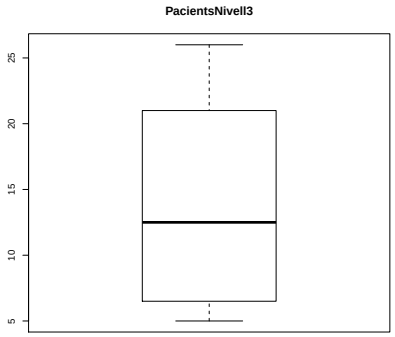
Els valors extrems són observacions amb característiques diferents a la resta, que tenen valors extremadament alts o extremadament baixos respecte al conjunt d'observacions analitzades. La mida de la mostra influeix directament en la probabilitat de que aquests valors apareguin. A major mida de la mostra major probabilitat de que aparegui algun valor extrem. La nostra mostra és suficientment petita com per no esperar cap valor extrem.

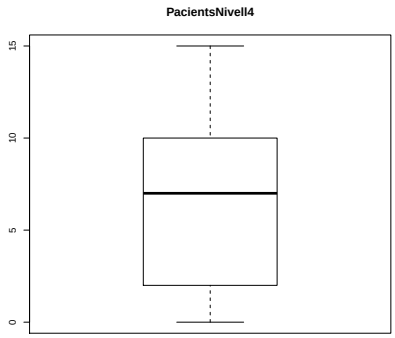
Per comprovar si les variables quantitatives presenten valors extrems començarem per representar en *boxplots* cadascuna de les variables.

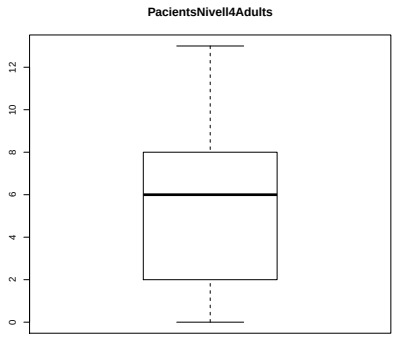
```
par(mfrow=c(3,3))
for(i in 1:ncol(bd)){if(is.integer(bd[,i])){boxplot(bd[,i],main=colnames(bd)[i])}}
```

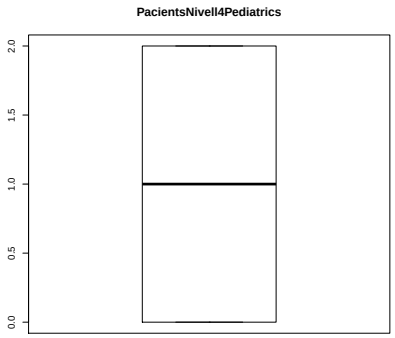

PacientsNivell1
1.0
0.5
0.0
−0.5
−1.0

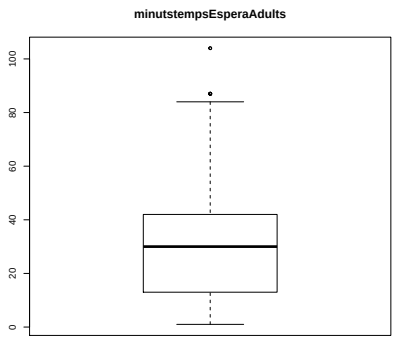

PacientsNivell2
6
5
4
3
2

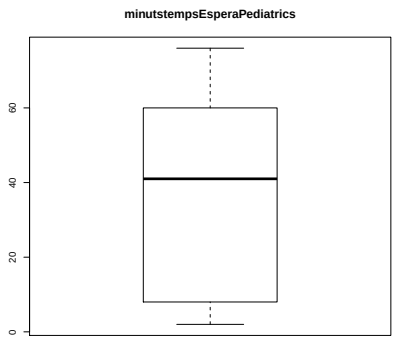

PacientsNivell3
25
20
15
10
5


PacientsNivell4
15
10
5
0


PacientsNivell4Adults
12
10
8
6
4
2
0


PacientsNivell4Pediatrics
2.0
1.5
1.0
0.5
0.0


minutstempsEsperaAdults
100
80
60
40
20
0


minutstempsEsperaPediatrics
60
40
20
0

El boxplot de la variable minutstempsEsperaAdults ens informa de la presència de valors extrems. Una altre manera de localitzar valors extrems i identificar-los és mitjançant la funció *boxplot.stats* amb l'argument *out*. Aquesta funció ens mostrarà els valors de qualsevol punt de dades que està més enllà dels extrems del bigotis dels boxplots.

```
for(i in 1:ncol(bd)){if(is.integer(bd[,i])){ print(cat(names(bd[i]),":",
boxplot.stats(bd[,i])$out,"\n"))}}
```

```
## PacientsNivell1 :
## NULL
## PacientsNivell2 :
## NULL
## PacientsNivell3 :
## NULL
## PacientsNivell4 :
## NULL
## PacientsNivell4Adults :
## NULL
## PacientsNivell4Pediatrics :
## NULL
## minutstempsEsperaAdults : 104 104 87 87 87 87 87 87 87
## NULL
## minutstempsEsperaPediatrics :
## NULL
```

La identificació dels valors extrems mitjançant la funció *boxplot.stats* ens alerta que no son dos punts els considerats com a valor extrem, i que els visualitzaven al boxplot, si no que són 9 files diferents que aporten en concret dos valors diferents. Tot i que quantitativament aquests valors corresponen a valors extrems en el domini de la variable, considerarem que representen una realitat de les dades del moment de captura al que pertanyen i no seran eliminats.

```
outliers=subset(bd,minutstempsEsperaAdults>=87)
outliers[,c(1,6,8)]
```

```
##     dataActualitzacio PacientsNivell4Adults horaActualitzacio
## 11         2018-04-11                     8             22:06
## 12         2018-04-11                     8             22:11
## 189        2018-04-12                     8             12:56
## 190        2018-04-12                     6             13:01
## 191        2018-04-12                     6             13:06
## 192        2018-04-12                     6             13:11
## 193        2018-04-12                     6             13:16
## 194        2018-04-12                     6             13:21
## 195        2018-04-12                     6             13:26
```

Un cop realitzades les tasques de pre-processament de les dades i abans d'avançar en l'anàlisi exportem un nou dataset amb les variables finals.

```
write.csv(bd,"CSAP_clean.csv")
```

5 Anàlisi de les dades.

5.1 Selecció dels grups de dades.

Realitzarem una selecció de dos grups per dur a terme l'anàlisi i comparar els resultats obtinguts: un per pacients adults i un altre per pacients pediàtrics. Dins de cadascun d'aquests dos grups es realitzarà l'estudi analític per torns per intentar donar resposta a la pregunta plantejada en aquest exercici.

```
adults<-bd[,c("Torn","PacientsNivell4Adults","minutstempsEsperaAdults",
              "dataActualitzacioCompleta")]
adults<-subset(adults, adults$PacientsNivell4Adults!=0)
pediatrics<-bd[,c("Torn","PacientsNivell4Pediatrics","minutstempsEsperaPediatrics",
                  "dataActualitzacioCompleta")]
pediatrics<-subset(pediatrics, pediatrics$PacientsNivell4Pediatrics!=0)
kable(head(adults), align = "c", caption = "Adults")
```

Taula 7: Adults

Torn	PacientsNivell4Adults	minutstempsEsperaAdults	dataActualitzacioCompleta
Tarda	9	84	2018-04-11 21:16:00
Tarda	10	84	2018-04-11 21:21:00
Tarda	8	51	2018-04-11 21:26:00
Tarda	8	74	2018-04-11 21:31:00
Tarda	10	74	2018-04-11 21:36:00
Tarda	11	74	2018-04-11 21:41:00

```
kable(head(pediatrics), align = "c",caption = "Pediàtrics")
```

Taula 8: Pediàtrics

Torn	PacientsNivell4Pediatrics	minutstempsEsperaPediatrics	dataActualitzacioCompleta
Tarda	2	60	2018-04-11 21:16:00
Tarda	2	60	2018-04-11 21:21:00
Tarda	1	75	2018-04-11 21:26:00
Tarda	2	75	2018-04-11 21:31:00
Tarda	2	75	2018-04-11 21:36:00
Tarda	2	75	2018-04-11 21:41:00

Realitzem un resum estadístic sobre els temps d'espera per torn per a cada grup:

```
by(adults$minutstempsEsperaAdults, adults$Torn, summary)
```

```
## adults$Torn: Mati
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   30.00   31.00   41.71   57.00   87.00
## ------------------------------------------------------------
```

```
## adults$Torn: Nit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    4.00   26.00   23.26   35.00   75.00
## ------------------------------------------------------------
## adults$Torn: Tarda
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    5.00   21.00   27.56   44.00  104.00
```

```
by(pediatrics$minutstempsEsperaPediatrics, pediatrics$Torn, summary)
```

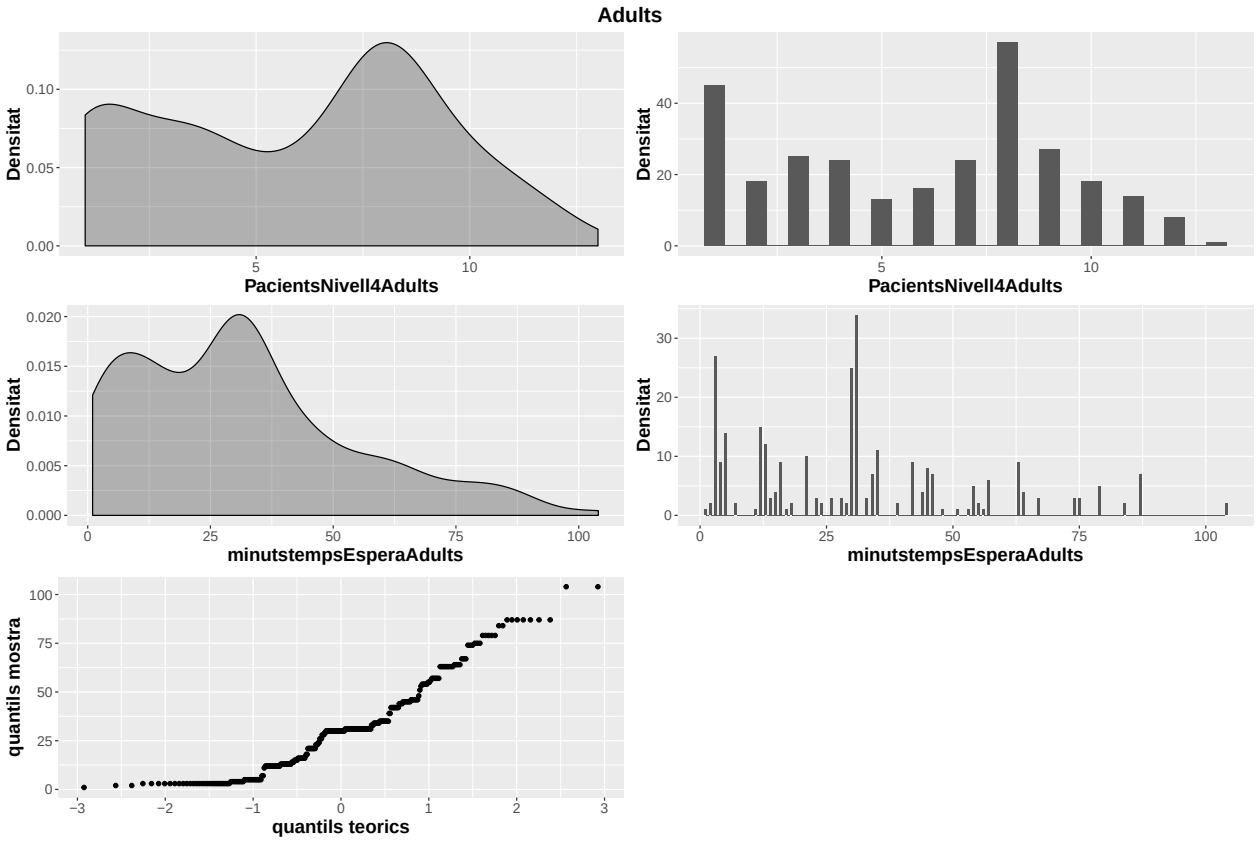
```
## pediatrics$Torn: Mati
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00    8.00    8.00   25.28   60.00   60.00
## ------------------------------------------------------------
## pediatrics$Torn: Nit
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   12.00   38.00   34.39   51.50   76.00
## ------------------------------------------------------------
## pediatrics$Torn: Tarda
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0     4.0     7.0    23.6    41.0    76.0
```

5.2 Comprovació de la normalitat i homogeneïtat de la variància.

Comprovarem la normalitat de les variables d'ambdos grups de dades. Comencem per visualitzat la distribució dels valors de les variables. Graficarem per a cada variable de cada grup de dades la seva gràfica de densitati l'histograma. Una altre manera de representar gràficament la normalitat de les mostres és mitjançant la funció *qqnorm*.

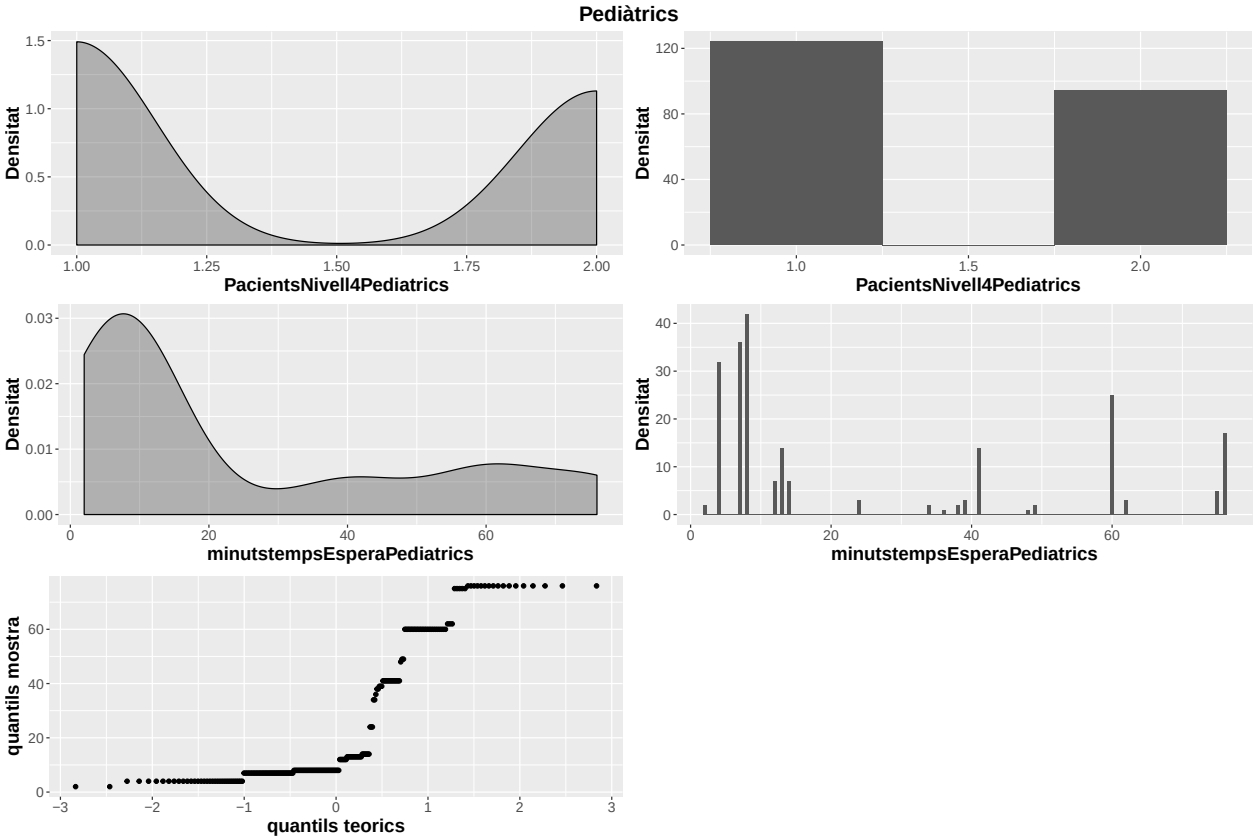
```
#adults
a1<-ggplot(adults, aes(x=PacientsNivell4Adults),
xlab = "Número pacients nivell 4 i 5 adults")+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
a2<-ggplot(adults, aes(x=PacientsNivell4Adults),
xlab = "Número pacients nivell 4 i 5 adults")+ylab("Densitat") +
  geom_histogram(binwidth=.5)+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
a3<-ggplot(adults, aes(x=minutstempsEsperaAdults),
xlab = "minuts espera adults")+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
a4<-ggplot(adults, aes(x=minutstempsEsperaAdults),
xlab = "minuts espera adults")+ylab("Densitat") +
  geom_histogram(binwidth=.5)+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
a9<-ggplot(adults, aes(sample=minutstempsEsperaAdults),
xlab = "minuts espera adults")+xlab("quantils teorics")+ylab("quantils mostra") +
```

```
  stat_qq()+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))

grid.arrange(a1, a2,a3,a4,a9, nrow = 3, ncol=2, top=textGrob("Adults",
gp=gpar(fontsize=18,font=2)))
```




```
#pediatrics
p1<-ggplot(pediatrics, aes(x=PacientsNivell4Pediatrics),
xlab = "Número pacients nivell 4 i 5 pediatrics")+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
p2<-ggplot(pediatrics, aes(x=PacientsNivell4Pediatrics),
xlab = "Número pacients nivell 4 i 5 pediatrics")+ylab("Densitat") +
  geom_histogram(binwidth=.5)+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
p3<-ggplot(pediatrics, aes(x=minutstempsEsperaPediatrics),
xlab = "minuts espera pediatrics")+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))
p4<-ggplot(pediatrics, aes(x=minutstempsEsperaPediatrics),
xlab = "minuts espera pediatrics")+ylab("Densitat")+geom_histogram(binwidth=.5)+
  theme(axis.text=element_text(size=12),axis.title=element_text(size=16,face="bold"))
p9<-ggplot(pediatrics, aes(sample=minutstempsEsperaPediatrics),
xlab = "minuts espera pediatrics")+xlab("quantils teorics")+ylab("quantils mostra") +
```

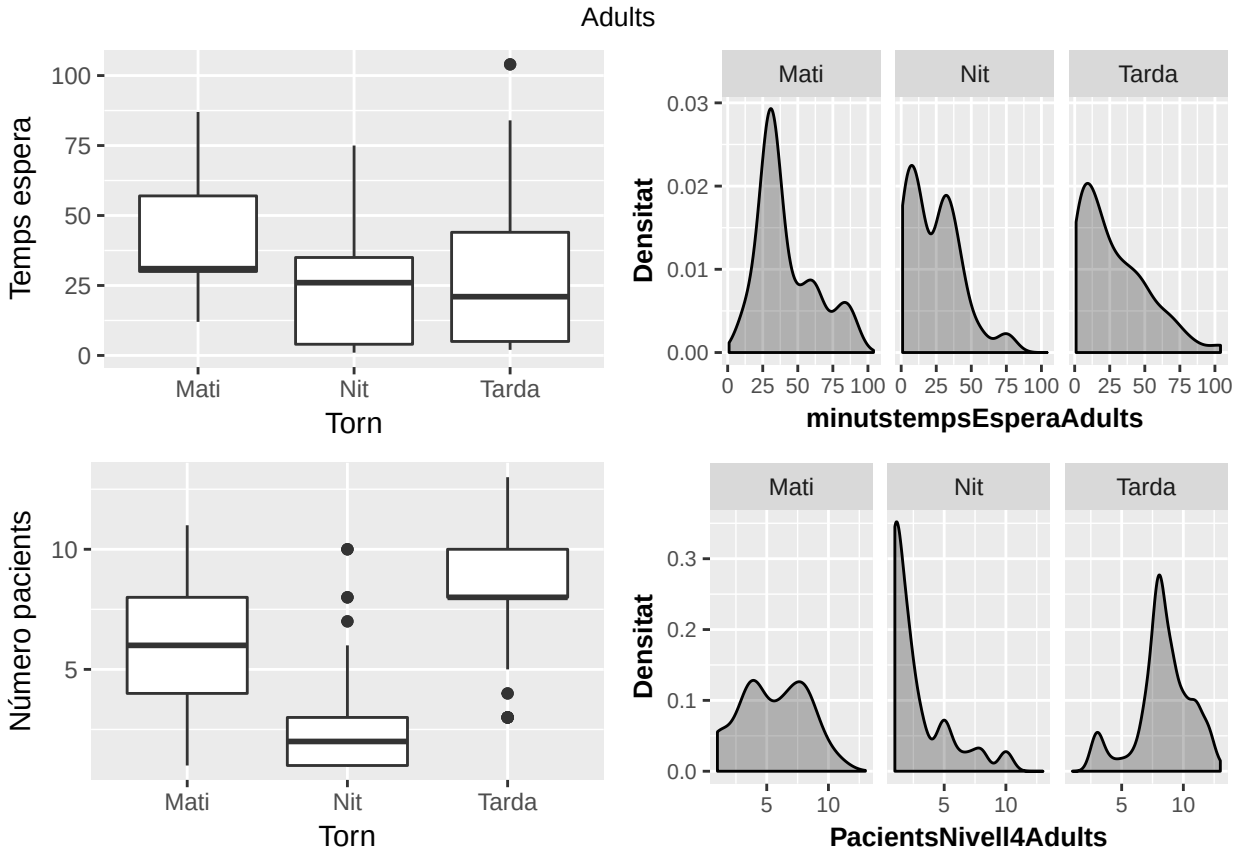
```
  stat_qq()+theme(axis.text=element_text(size=12),
        axis.title=element_text(size=16,face="bold"))

grid.arrange(p1, p2,p3, p4,p9,nrow = 3, ncol=2, top=textGrob("Pediàtrics",
gp=gpar(fontsize=18,font=2)))
```



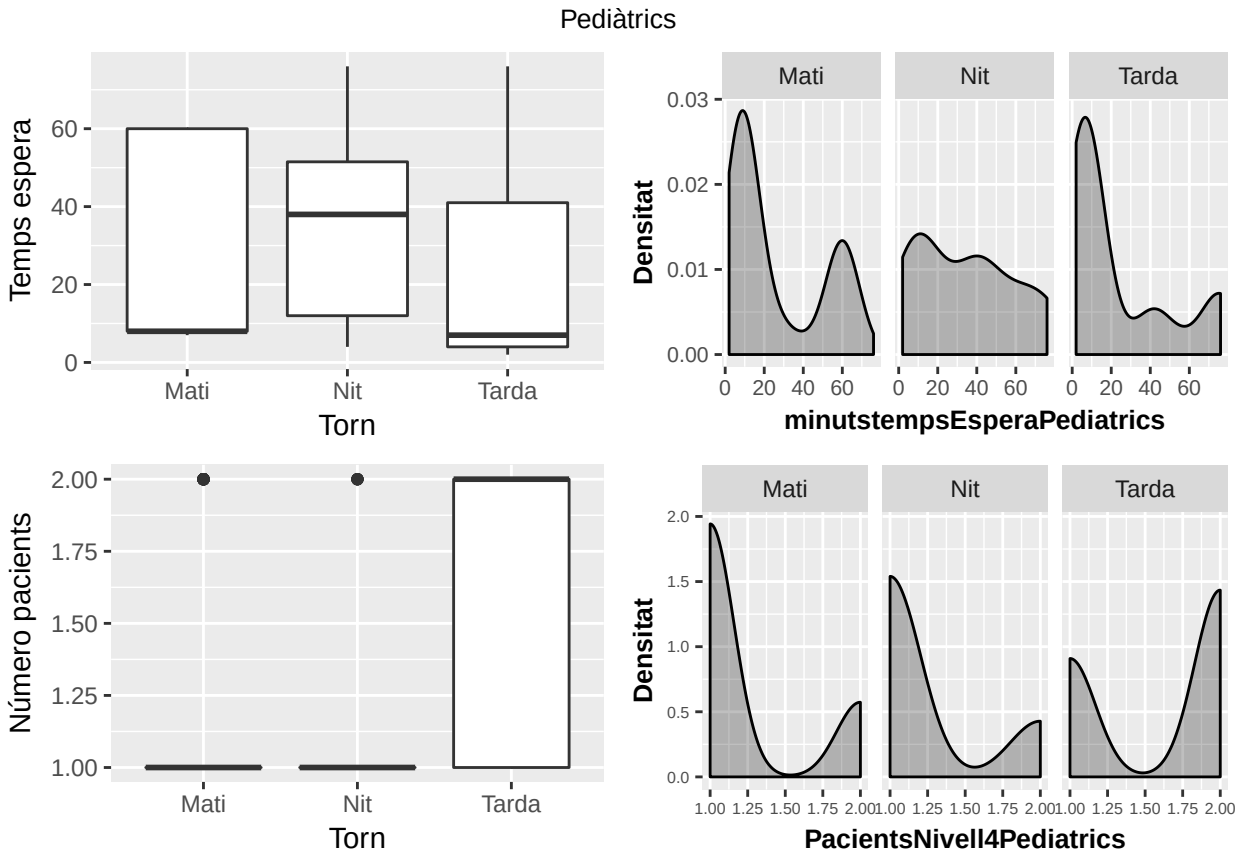
Les representacions gràfiques denoten que cap de les variables segueix una distribució normal.

Analitzarem les variables per ambdos grups per torn.

```
a5<-ggplot(adults, aes(x=Torn, y=adults$minutstempsEsperaAdults))+
  ylab("Temps espera") +geom_boxplot()
a6<-ggplot(adults, aes(x=minutstempsEsperaAdults))+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=8),
        axis.title=element_text(size=10,face="bold")) + facet_wrap(~Torn)
a7<-ggplot(adults, aes(x=Torn, y=adults$PacientsNivell4Adults))+
  ylab("Número pacients") +geom_boxplot()
a8<-ggplot(adults, aes(x=PacientsNivell4Adults))+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=8),
        axis.title=element_text(size=10,face="bold")) + facet_wrap(~Torn)
grid.arrange(a5, a6,a7,a8, nrow = 2, ncol=2, top=textGrob("Adults",
gp=gpar(fontsize=10)))
```




```
p5<-ggplot(pediatrics, aes(x=Torn, y=pediatrics$minutstempsEsperaPediatrics))+
  ylab("Temps espera") +geom_boxplot()
p6<-ggplot(pediatrics, aes(x=minutstempsEsperaPediatrics))+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=8),
        axis.title=element_text(size=10,face="bold")) + facet_wrap(~Torn)
p7<-ggplot(pediatrics, aes(x=Torn, y=pediatrics$PacientsNivell4Pediatrics))+
  ylab("Número pacients") +geom_boxplot()
p8<-ggplot(pediatrics, aes(x=PacientsNivell4Pediatrics))+ylab("Densitat") +
  geom_density(fill="black", alpha=0.25)+theme(axis.text=element_text(size=6),
        axis.title=element_text(size=10,face="bold")) + facet_wrap(~Torn)
grid.arrange(p5, p6,p7,p8, nrow = 2, ncol=2, top=textGrob("Pediàtrics",
gp=gpar(fontsize=10)))
```



Els boxplots en adults ens mostren que encara que la mitjana de pacients sigui major en el torn de tarda, el temps mig d'espera és superior al torn de matí. Al grup de pediàtrics, tot i que l'afluència de nens és molt petita en cada torn, el temps mig d'espera destaca en el torn de nit.

Continuarem amb l'anàlisi i realitzarem un test de Anderson-Darling[2] sobre cada variable per confirmar que no segueixen una distribució normal. El test estableix com a hipòtesi nul · la que la població està distribuïda normalment. Si el p-valor es menor a un nivell de significació de 0.05 aleshores la hipòtesi nul · la es rebutjada.

```
print(ad.test(adults$PacientsNivell4Adults))
```

```
##
##  Anderson-Darling normality test
##
## data:  adults$PacientsNivell4Adults
## A = 7.1788, p-value < 2.2e-16
```

```
print(ad.test(adults$minutstempsEsperaAdults))
```

```
##
##  Anderson-Darling normality test
##
## data:  adults$minutstempsEsperaAdults
## A = 6.147, p-value = 3.873e-15
```

[2] https://en.wikipedia.org/wiki/Anderson\T1\textendashDarling_test

```r
print(ad.test(pediatrics$PacientsNivell4Pediatrics))
```

```
##
##  Anderson-Darling normality test
##
## data:  pediatrics$PacientsNivell4Pediatrics
## A = 39.99, p-value < 2.2e-16
```

```r
print(ad.test(pediatrics$minutstempsEsperaPediatrics))
```

```
##
##  Anderson-Darling normality test
##
## data:  pediatrics$minutstempsEsperaPediatrics
## A = 23.494, p-value < 2.2e-16
```

Els resultats del test confirmen que cap de les variables segueix una distribució normal, donat que el p-valor és significativament inferior a 0.05.

Les mostres que disposem superen el valor de n>3 i pel Teorema del Límit Central s'estableix que la distribució de la mitjana mostral segueix aproximadament una normal.

Per determinar si existeixen diferències significativas entre els diferents torns realizarem un análisis de variància Realitzem una comprovació de l'homogeneïtat de la variància de les dos poblacions mitjançant el test de Fligner-Killeen, el qual resulta una millor opció quan les dades no es distribueixen de forma normal. La hipòtesi nul · la del test estableix que les variances son homogènees i l'alternativa que són heteroscedàstiques.

```r
fligner.test(minutstempsEsperaAdults~interaction(Torn),data=adults)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  minutstempsEsperaAdults by interaction(Torn)
## Fligner-Killeen:med chi-squared = 5.5441, df = 2, p-value =
## 0.06253
```

```r
fligner.test(minutstempsEsperaPediatrics~interaction(Torn),data=pediatrics)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  minutstempsEsperaPediatrics by interaction(Torn)
## Fligner-Killeen:med chi-squared = 3.6446, df = 2, p-value = 0.1617
```

Els resultats dels tests per amdos poblacions són significatius essent en ambdos casos el p-valor superior a 0,05, aceptant-se doncs la hipòtesi nul · la d'homogeneitat en les variances dels torns.

5.3 Aplicació de proves estadístiques per comparar els grups de dades.

Realitzarem un anàlisi de la variància (ANOVA) sobre la mitjana de la variable *minutstempsEsperaAdults* del grups de dades *adults* per cada torn, per comprovar si existeix alguna relació del temps d'espera amnb el torn o si pel contrari la mitjana del temps d'espera és la mateixa indiferentment del torn. Establim la hipòtesi nul · la i l'alternativa següents:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

H_0 :No hi ha diferència entre les mitjanes de temps d'espera de cada torn

H_1 : Al menys un par de mitjanes són significativament diferents entre elles

```
ANOVA= aov( adults$minutstempsEsperaAdults ~ adults$Torn)
summary(ANOVA)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## adults$Torn   2  16999    8500   18.34 3.19e-08 ***
## Residuals   287 133003     463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Donat que el p-value és significativament inferior a 0,05 existeixen evidències estadístiques suficients com per a considerar que almenys dos mitjanes són diferents.

Per determinar la intensitat de la relació entre els torns calcularem l'estadístic η^2 que ens mesura quina part de la variació del temps d'espera és atribuible a la variable Torn.

```
eta_quadrat <- 16999/(16999 + 133003)
eta_quadrat
```

```
## [1] 0.1133252
```

Realitzem un test amb el mètode HSD de Tukey per identificar quins torns tenen certa relació respecte a les mitjanes de temps d'espera.

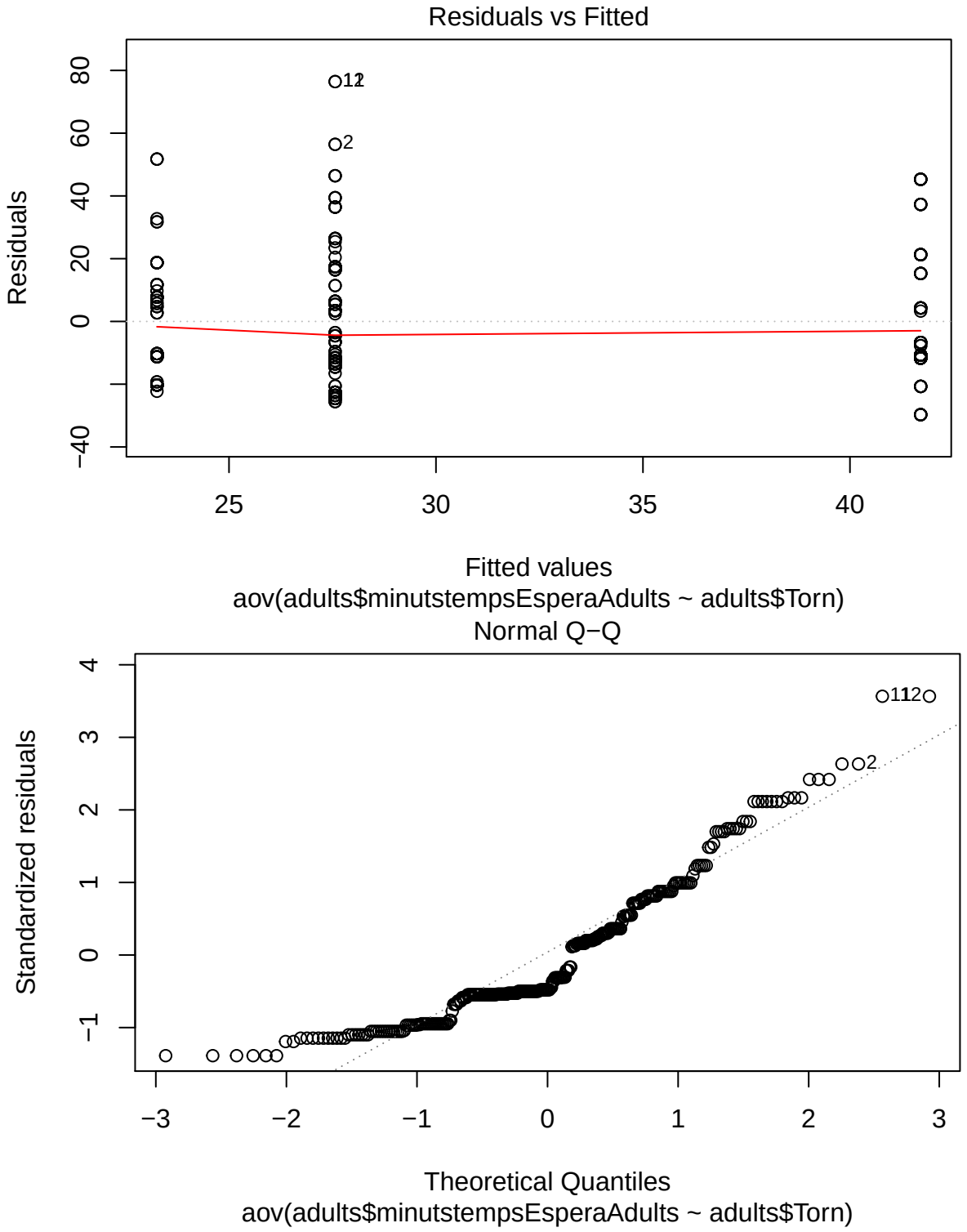
```
TukeyHSD(ANOVA)
```

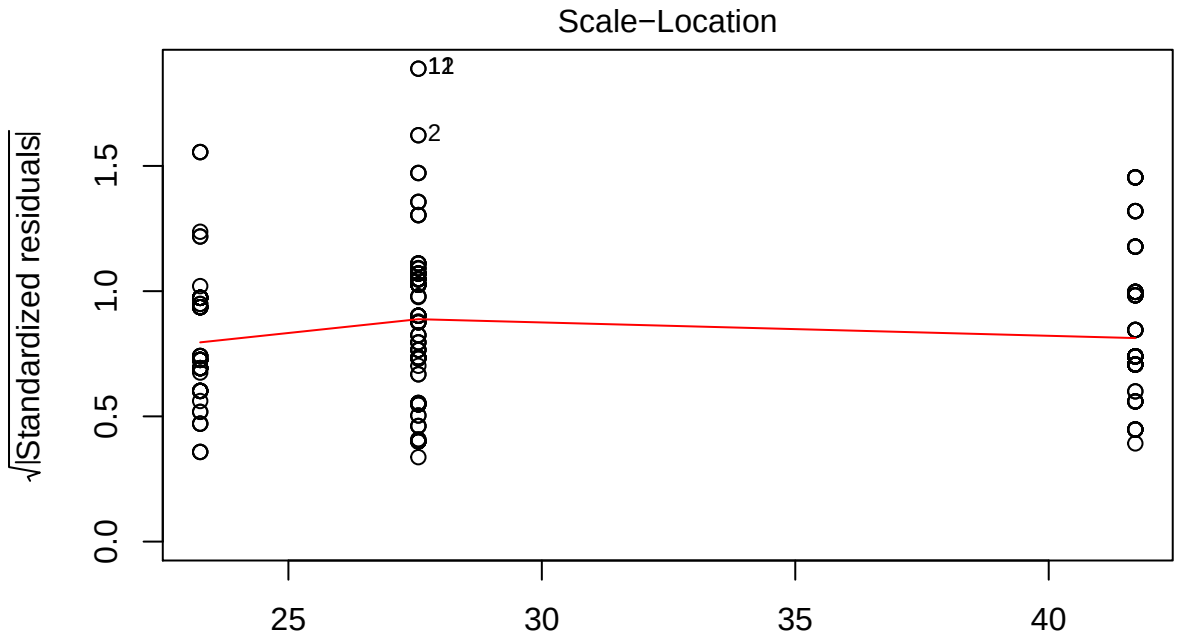
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = adults$minutstempsEsperaAdults ~ adults$Torn)
##
## $`adults$Torn`
##                  diff        lwr        upr     p adj
## Nit-Mati   -18.448593 -26.207427 -10.689759 0.0000001
## Tarda-Mati -14.144231 -21.128415  -7.160047 0.0000087
## Tarda-Nit    4.304362  -3.138106  11.746830 0.3620205
```

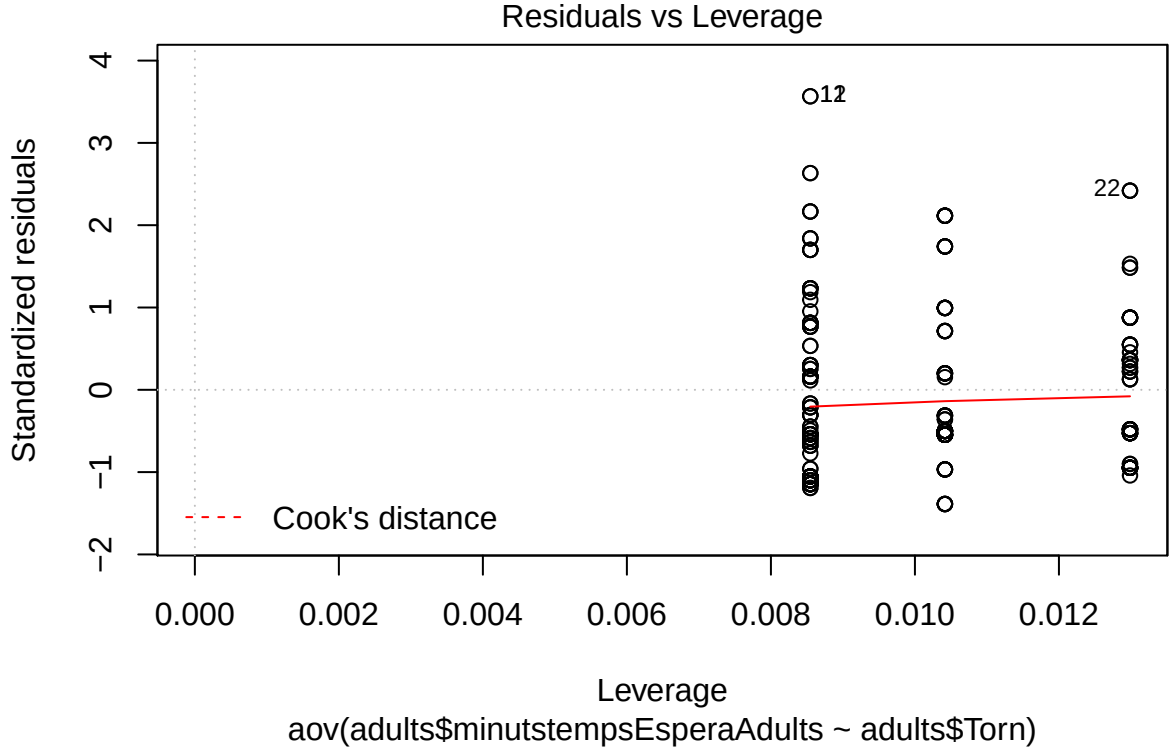
6 Representació dels resultats a partir de taules i gràfiques.

Grafiquem els resultats de l'anàlisi ANOVA com del test de Tukey.

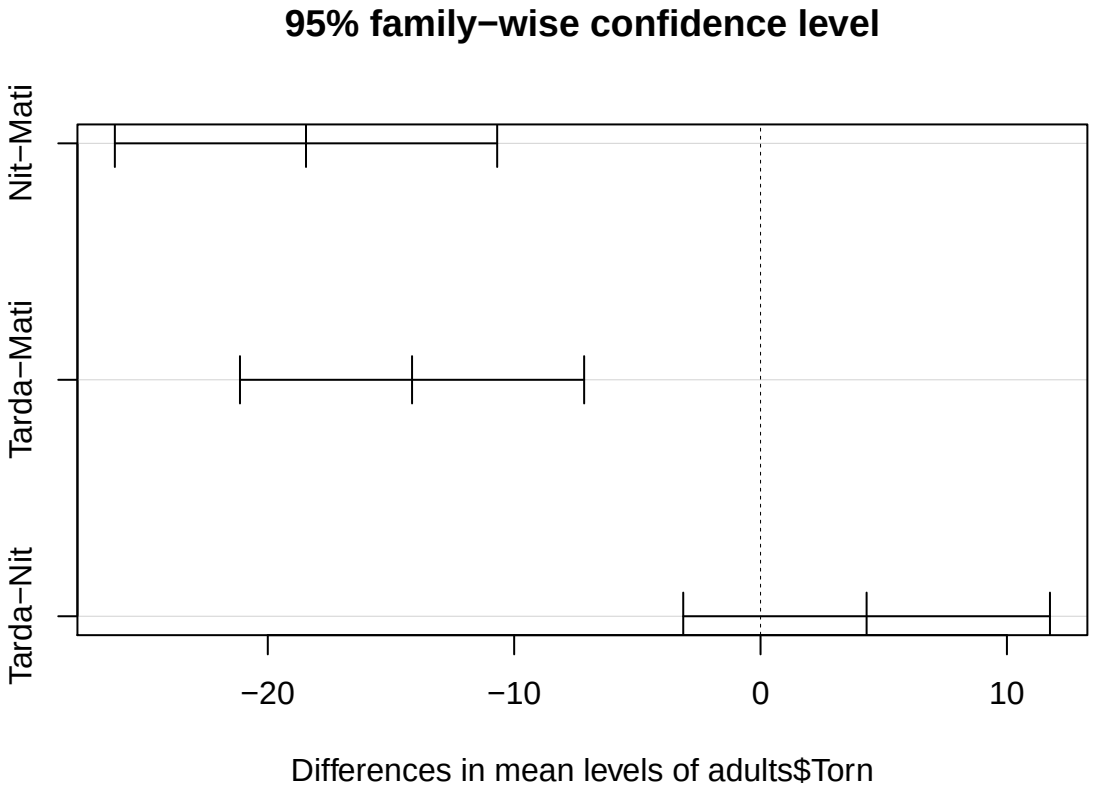
```
plot(ANOVA)
```








```
plot(TukeyHSD(ANOVA))
```



7 Resolució del problema. Conclusions.

En aquest exercici partiem d'un dataset aconseguit mitjançant web-scraping amb dades de temps d'espera dels pacients pendents de ser atesos en un departament d'urgències d'un centre sanitari. Plantejavem una sèrie de suposicions que, quan assistim a un servei d'urgències donem per preconcebides. Exposavem que és molt habitual pensar que el temps d'espera que ens toca passar en la sala d'espera és raonable pensar que ha de d'estar condicionat pel volum de pacients que podem visualitzar que també esperen a la sala. Però hem plantejat una pregunta intentant obviar aquesta relació i enfocant-la a analitzar si el torn horari, és a dir, si el fet de que sigui el torn de mati, tarda o nit podria tenir relació amb la mitjana del temps d'espera, sempre en els pacients que no passen directament a ser atesos: els pacients classificat com nivell 4 o 5 d'urgència. Els anàlisis gràfics realitzats al principi de l'exercici ja ens encoratjaven a continuar amb l'estudi. Es podia contrastar com el volum de pacients en els torns no s'acabava de correspondre amb la distribució dels temps d'espera als mateixos torns. He decidit centrar l'estudi en el grup d'adults i realitzar un constrast d'hipòtesi sota l'anàlisi de la variància els resultats del qual ens concluen que, existeix significància estadística suficient que denota que els temps d'espera no segueixen la mateixa distribució entre els torns. Els resultats d'ANOVA ens informa que existeixen diferències significatives entre el temps d'espera per adults entre el torn de matí i el torn de tarda, i entre el torn de matí i el torn de nit. No detecta diferències significatives entre el torn de tarda i el de nit. Per tant l'anàlisi dona resposta a la pregunta plantejada i ens fa conclure que si existeixen relació entre el temps d'espera per als pacients adults dels nivells 4 i 5 i el torn del dia en el que esperem a ser atesos.

8 Codi.

S'annexa tot el codi implementat per realizar l'anàlisi sencer i de forma continua.

9 Referències

R-Documentation

Cran R-Project

Cookbook for R

http://www.csap.cat/memoria/2016/activitat_2016/urgencies.html

https://en.wikipedia.org/wiki/Anderson-Darling_test

http://wiki.stat.ucla.edu/socr/index.php/AP_Statistics_Curriculum_2007_NonParam_VarIndep