PRAC2 - Neteja i validació de les dades

Rubén Vasallo González

26 de December 2018

Contents

Descripción del dataset

El conjunt de dades que anem a analitzar s'ha extret de la competició de la web de Kaggle anomenada "Titanic: Machine Learning from Disaster" (https://www.kaggle.com/c/titanic) que permet iniciar-se en les competicions que aquesta pagina web te. Aquest conjunt de dades (o dataset) te 891 observacions que representen algunes de les característiques dels passatgers, junt amb un atribut que indica si aquest van sobreviure al famós enfonsament o no. A continuació detallem el significat de cada atribut i, si es te, els possibles valors que poden tenir (en cas de ser atributs qualitatius)

Variable	Definició	Key
PassengerId	Identificador del passatger	
Survived	Sobreviu	0 = No, 1 = Yes
pclass	Classe de Tiquet	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Nom del passatger	
Sex	Sexe	male, female
Age	Edat en anys	
SibSp	# de germans / esposes a bord del Titanic	
Parch	# de pares / fills a bord del Titanic	
Ticket	Numero de Tiquet	
Fare	Tarifa del passatger	
Cabin	Numero de Cabina	
Embarked	Port d'embarcament	C = Cherbourg, Q = Queenstown, S = Southampton

Quina pregunta/problema pretén respondre?

L'enfonsament del Titanic va ser un dels naufragis mes famós de l'historia. A dia d'avui encara presenta grans misteris i preguntes sense resolució. Una de les raons per les quals el naufragi va comportar aquesta pèrdua de vida va ser que no hi havia prou barques salvavides per als passatgers i la tripulació. Ates a la norma moral de que en cas d'enfonsament, les dones i els nens haurien estat els primers en ser evacuats (https://es.wikipedia.org/wiki/Mujeres_y_ni%C3%B1os_primero), a l'hora de la veritat, això no va ser així, i tot i que hi va haver algun element de sort en la supervivència de l'enfonsament, la historia ens ha demostrat que alguns grups de persones tenien més probabilitats de sobreviure que altres, com ara la classe alta.

L'objectiu d'aquest dataset es el estudi de les característiques dels passatgers que van sobreviure per tractar de crear un model d'aprenentatge automàtic que sigui capaç de predir si nous passatgers amb característiques

semblants sobreviurien o no.

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

Per assolir l'objectiu primer carregarem el dataset i li farem una ullada als 6 primers resultats.

```
# Carregar el fitxer de dades en R
titanicdataset <- read.csv2(file = "train.csv", header = TRUE, sep = ",", quote = "\"", fill = TRUE, en
# breu resumen dels atributs
head(titanicdataset)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex  Age SibSp
## 1                             Braund, Mr. Owen Harris   male   22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female   38     1
## 3                              Heikkinen, Miss. Laina female   26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female   35     1
## 5                            Allen, Mr. William Henry   male   35     0
## 6                                    Moran, Mr. James   male <NA>     0
##   Parch            Ticket    Fare Cabin Embarked
## 1     0         A/5 21171    7.25  <NA>        S
## 2     0          PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282   7.925  <NA>        S
## 4     0            113803    53.1  C123        S
## 5     0            373450    8.05  <NA>        S
## 6     0            330877  8.4583  <NA>        Q
```

Veiem que tenim 12 atributs i que no tots tenen valors assignats. Cridem a la funció summary per veure un resum dels valors del dataset sencer.

```
summary(titanicdataset)
```

```
##   PassengerId       Survived          Pclass
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000
##  Median :446.0   Median :0.0000   Median :3.000
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##                                     Name         Sex           Age
##  Abbing, Mr. Anthony                   :  1   female:314   24     : 30
##  Abbott, Mr. Rossmore Edward           :  1   male  :577   22     : 27
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1                18     : 26
##  Abelson, Mr. Samuel                   :  1                19     : 25
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                28     : 25
##  Adahl, Mr. Mauritz Nils Martin        :  1                (Other):581
##  (Other)                               :885                NA's   :177
##      SibSp           Parch             Ticket         Fare
##  Min.   :0.000   Min.   :0.0000   1601    :  7   8.05   : 43
```

```
##  1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   13     : 42
##  Median :0.000   Median :0.0000   CA. 2343:  7   7.8958 : 38
##  Mean   :0.523   Mean   :0.3816   3101295 :  6   7.75   : 34
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   26     : 31
##  Max.   :8.000   Max.   :6.0000   CA 2144 :  6   10.5   : 24
##                                   (Other) :852   (Other):679
##          Cabin     Embarked
##  B96 B98    :  4   C   :168
##  C23 C25 C27:  4   Q   : 77
##  G6         :  4   S   :644
##  C22 C26    :  3   NA's:  2
##  D          :  3
##  (Other)    :186
##  NA's       :687
```

De tots els atributs, veiem que els de identificador del passatger, nom del passatger i numero de tiquet son irrellevants a l'hora de predir si un passatger sobreviu o no, per el que els descartem del dataset. Ens quedem amb la resta d'atributs que a priori, podrien ser rellevants a l'hora de predir la supervivència d'un passatger.

```
titanicdataset$PassengerId <- NULL
titanicdataset$Name <- NULL
titanicdataset$Ticket <- NULL
```

També podem veure que el programa R ha detectat gaire be tots els atributs com a numèrics (Excepte l'atribut sexe i edat). Realment en aquest dataset tots els valors es poden considerar com a categòrics ja que la informació a la que fa referencia està acotat (no es infinit) a excepció de l'atribut edat i taxa (Age, Fare) que poden tenir sentit que siguin contínues de cara a detectar possibles valors outliers o erronis.

```
sapply(titanicdataset, function(x) class(x))
```

```
##  Survived    Pclass       Sex       Age     SibSp     Parch      Fare
## "integer" "integer"  "factor"  "factor" "integer" "integer"  "factor"
##     Cabin  Embarked
##  "factor"  "factor"
```

Transformem els atributs a categòrics.

```
# Transformem els atributs necessaris a categòrics.
titanicdataset[,'Survived']<-factor(titanicdataset[,'Survived'])
titanicdataset[,'Pclass']<-factor(titanicdataset[,'Pclass'])
titanicdataset[,'SibSp']<-factor(titanicdataset[,'SibSp'])
titanicdataset[,'Parch']<-factor(titanicdataset[,'Parch'])

# Transformem els atributs necessaris a contínues
titanicdataset$Age <- as.numeric(titanicdataset$Age)

#options(digits=4)
titanicdataset$Fare <- as.character(titanicdataset$Fare)
titanicdataset$Fare <- as.numeric(titanicdataset$Fare)
```

Neteja de les dades.

Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Primer comprovem si tenim cap observació amb atributs a null.

```
sapply(titanicdataset, function(x) sum(is.na(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare    Cabin
##        0        0        0      177        0        0        0      687
## Embarked
##        2
```

En aquest cas tenim valors buits en els atributs Age, Cabin i Embarked

Analitzem cadascun d'ells.

Observem que tenim en el atribut Embarked valors per informar (Tenim 2 passatgers que no tenen port d'embarcament).

```
summary(titanicdataset$Embarked)
```

```
##    C    Q    S NA's
##  168   77  644    2
```

Ja que no tenim cap manera de recuperar la informació que ens falta intentarem aproximar el origen dels embarcaments de les observacions que no tenen aquesta informació utilitzant el mètode dels veïns mes propers (utilitzant la funció KNN del paquet VIM), que imputa el valor dels resultats (k veïns) mes propers. (Sempre serà millor aproximar el resultat que descartar-lo ja que tindrem un model amb menor marge d'error)

```
suppressWarnings(suppressMessages(library(VIM)))

titanicdataset$Embarked <- kNN(titanicdataset)$Embarked
```

Observem que tenim el mateix problema amb el atribut Age. Igual que ens passa en el cas aterior, no tenim manera de recuperar aquesta informació per el que intentarem Aproximar-lo mitjançant el mètode dels veïns mes propers.

```
suppressWarnings(suppressMessages(library(VIM)))

titanicdataset$Age <- kNN(titanicdataset)$Age
```

Finalment observem que el atribut Cabin li falta 687 valors. Aquest valor es mes complicat d'aproximar ja que no podem calcular quin era la cabina on viatjava els passatgers que no tenen una cabina assignada o si viatjaven sense cabina. En un principi vaig pensar que els passatgers que no tenien cabina assignada es que viatjaven al passadís però analitzant mes les dades, es pot observar que hi ha 40 passatgers de primera classe que no tenen una cabina assignada, i costa de creure que un passatger de primera viatges sense cabina.

```
table(titanicdataset[is.na(titanicdataset$Cabin),]$Pclass)
```

```
##
##   1   2   3
##  40 168 479
```

Descartem l'atribut.

```
titanicdataset$Cabin <- NULL
```

Comprovem que no tenim cap atribut amb valors buits.

```
sapply(titanicdataset, function(x) sum(is.na(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0        0        0        0        0        0
```

A continuació comprovem si tenim cap atribut amb valor a 0.

```
sapply(titanicdataset, function(x) sum(x == 0))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##      549        0        0        0      608      678       15        0
```

En aquest cas tenim valors a 0 en els atributs Survived, SibSp Parch i Fare

Analitzem cadascun d'ells.

- Survived te sentit que tingui valors a 0 ja que son tots aquells passatgers que no van sobreviure.
- SibSp te sentit que tingui valors a 0 ja que son tots aquells passatgers que no tenien cap germà o esposa a bord.
- Parch te sentit que tingui valors a 0 ja que son tots aquells passatgers que no tenien cap pare, mare o fill a bord.
- Fare te sentit que tingui valors a 0 ja que podrien ser tots aquells passatgers que van ser invitats, no van pagar cap tarifa per viatjar a bord o eren tripulants.

No modifiquem pas cap valor.

A continuació comprovem si tenim cap atribut amb valor string buit.

```
sapply(titanicdataset, function(x) sum(x == ""))
```

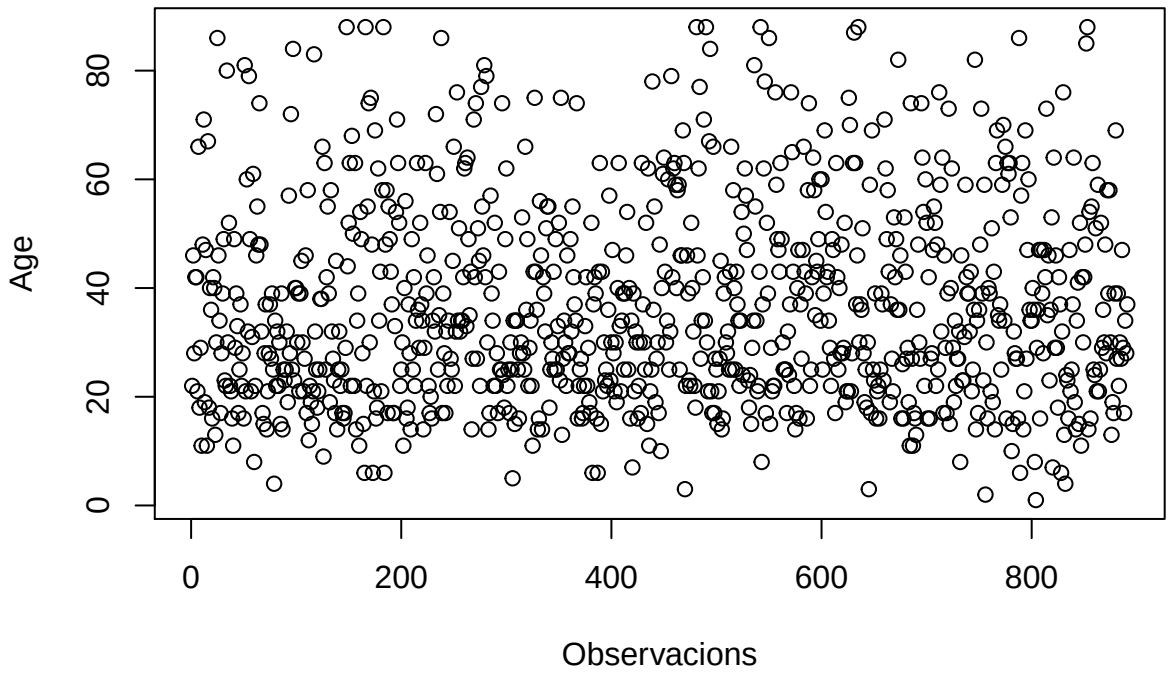
```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##        0        0        0        0        0        0        0        0
```

No tenim cap observació amb strings vuits.

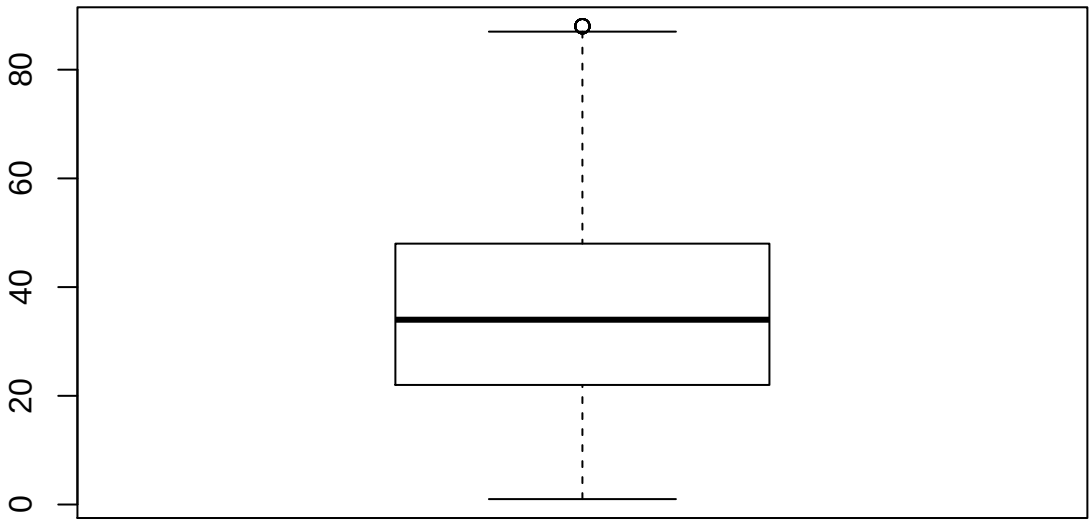
Identificació i tractament de valors extrems.

Com hem comentat anteriorment, la majoria dels atributs que te aquest dataset son categòrics (tenen valors finits) per el que es difícil detectar valors outliers. En aquest cas podríem intentar detectar outliers visualitzant en el atribut Edat, mostrant la distribució de les dades en gràfics.

```
plot(titanicdataset$Age, xlab="Observacions", ylab="Age")
```



```r
boxplot(as.numeric(titanicdataset$Age))
```



Podem veure que tenim outliers tant per un extrem com per altre. Això es normal ja que en el Titànic viatjaven tant nens com gent d'avançada edat.

Avaluem a continuació l'atribut Fare. Aparentment observem que te valors des de 0 fins a 512.3292. Tot i que el màxim es una xifra elevada, podria ser real ja que el Titanic tenia compartiments dedicats a la alta elit que podria permetre pagar un preu elevat.

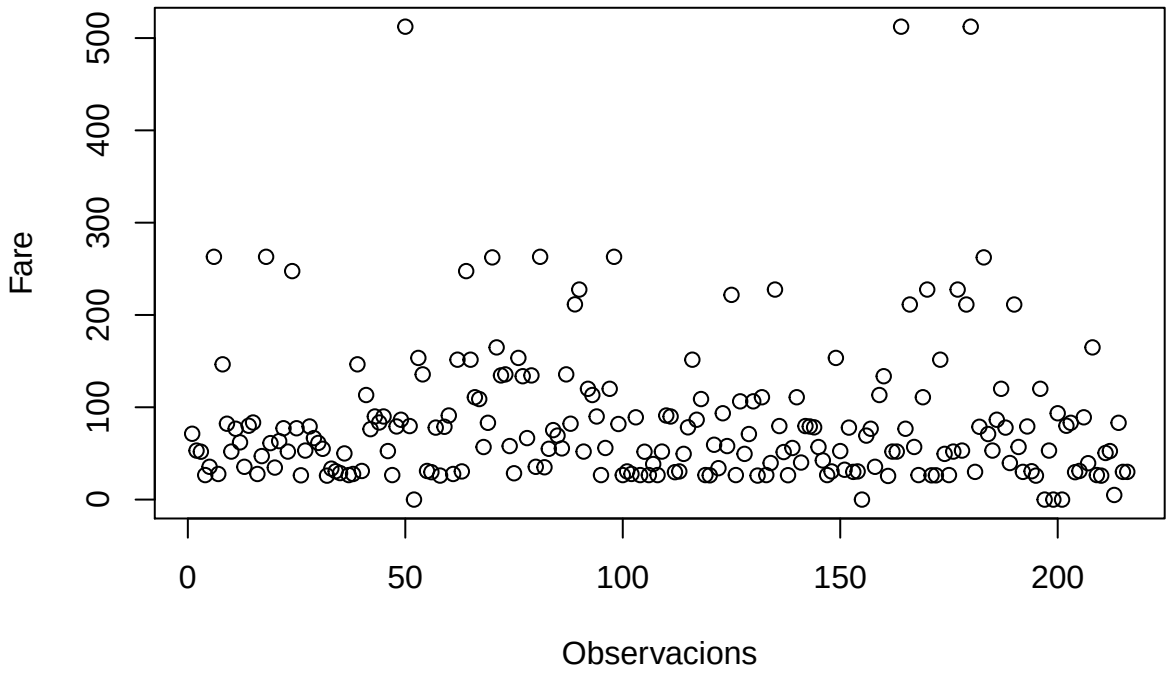
Tot i així agruparem els resultats per classe i avaluarem si hi ha cap valor que que sigui sospitós.

Comencem visualitzant els valors per a la primera classe.

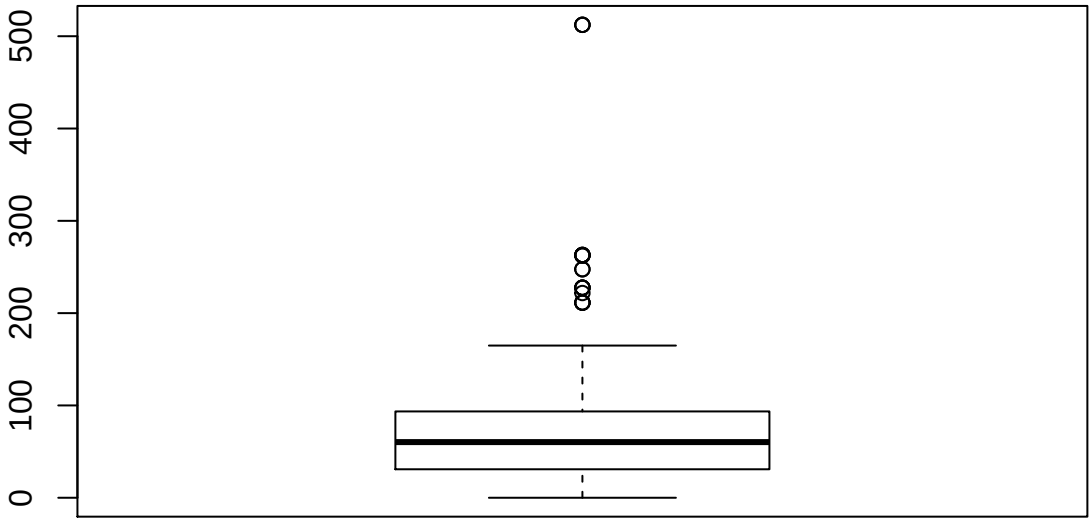
```r
passenger1class <- titanicdataset[titanicdataset$Pclass == "1",]
summary(passenger1class)
```

```
##  Survived Pclass      Sex           Age        SibSp   Parch
##  0: 80    1:216   female: 94   Min.   : 5.00   0:137   0:163
##  1:136    2:  0   male  :122   1st Qu.:32.00   1: 71   1: 31
##           3:  0                 Median :46.00   2:  5   2: 21
##                                 Mean   :46.42   3:  3   3:  0
##                                 3rd Qu.:60.50   4:  0   4:  1
##                                 Max.   :87.00   5:  0   5:  0
##                                                 8:  0   6:  0
##       Fare        Embarked
##  Min.   :  0.00   C: 86
##  1st Qu.: 30.92   Q:  2
##  Median : 60.29   S:128
##  Mean   : 84.15
##  3rd Qu.: 93.50
##  Max.   :512.33
##
```

```r
plot(passenger1class$Fare, xlab="Observacions", ylab="Fare")
```



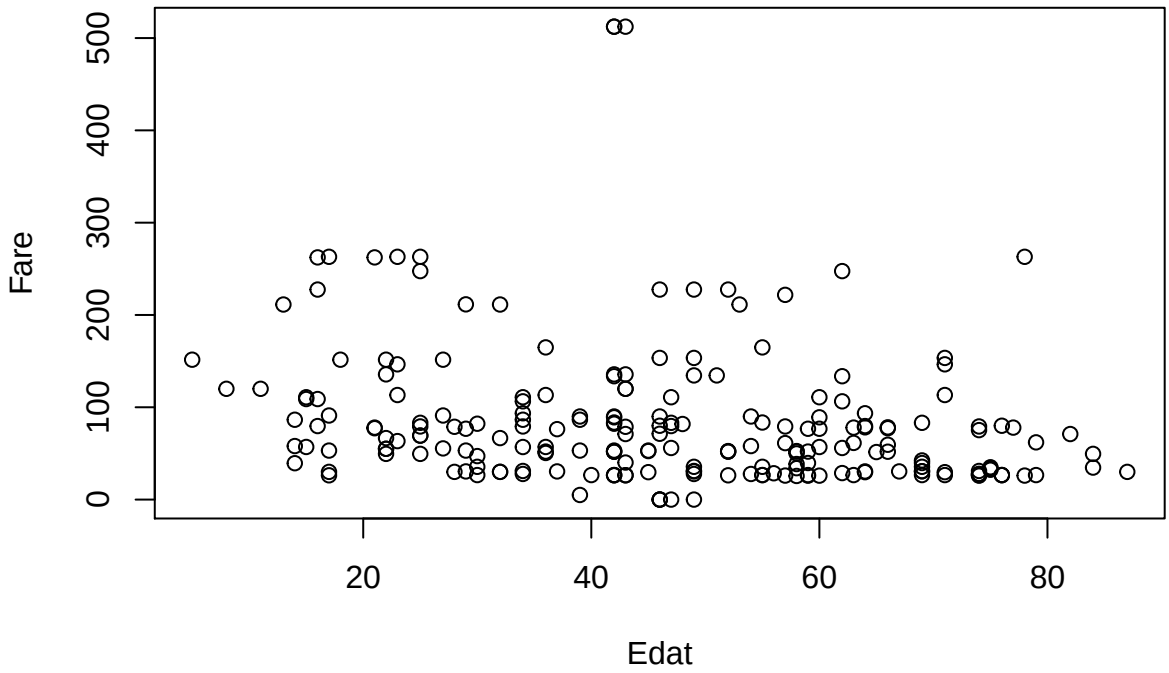

```
boxplot(passenger1class$Fare)
```



Veiem que tenim 3 valors molt alts per els passatgers de primera classe. Te tota la pinta que ha estat un error d'inserció de la informació. No tenim manera de recuperar el valor original i ja que son pocs valors, podríem descartar aquestes observacions. A mes a mes veiem que tenim valors molts baixos. Podem suposar que al ser un viatge inaugural, podria ser que alguns tiquets fossin regals a preu simbòlic per a les personalitats.

Tot i així fem primer un encreuament entre el preu i l'edat del client per acabar de contrastar els valors.

```
plot(passenger1class$Age, passenger1class$Fare,  xlab="Edat", ylab="Fare")
```



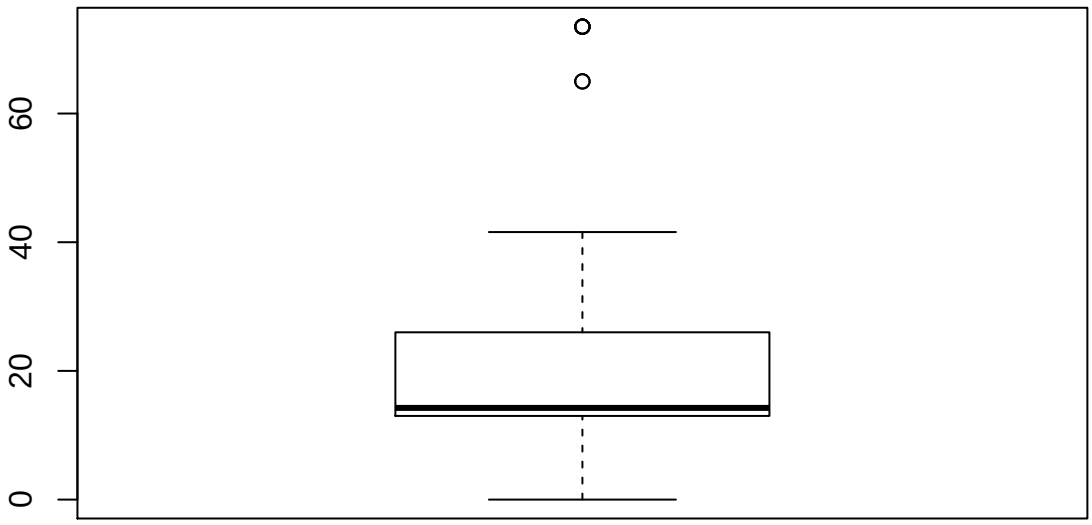
Definitivament no te sentit. Observem que hi han persones amb edat inferior a 20 anys que han pagat un preu elevat (entre 100 i 300).

Fem el mateix estudi amb els passatgers de 2 classe.

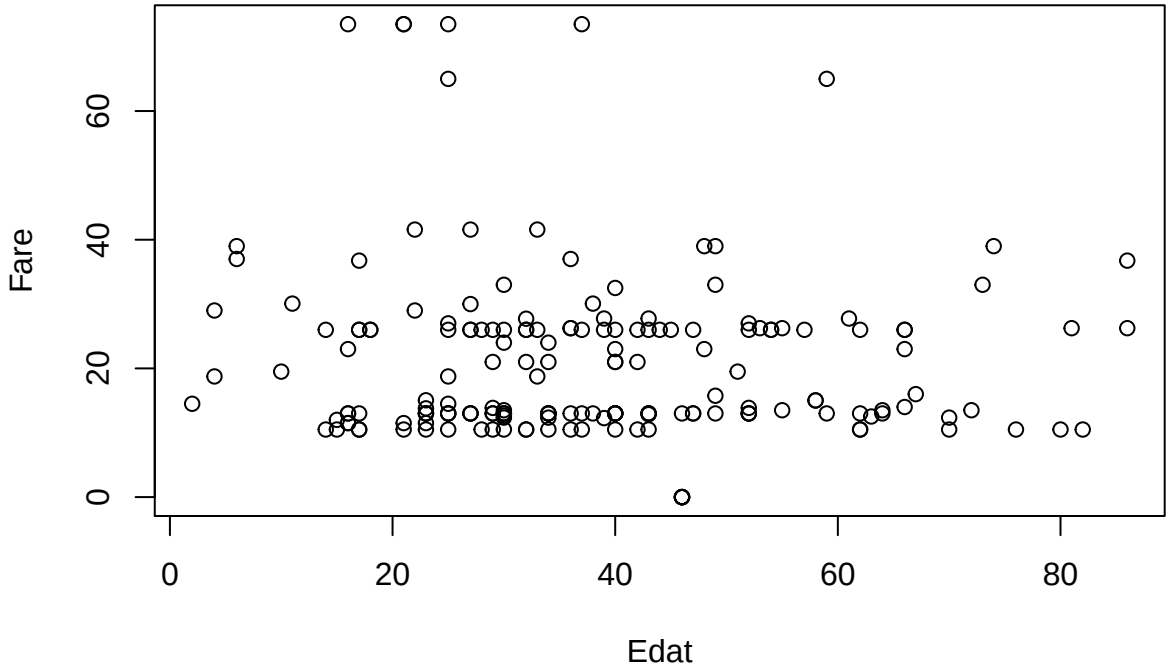
```
passenger2class <- titanicdataset[titanicdataset$Pclass == "2",]
summary(passenger2class)
```

```
##  Survived Pclass      Sex           Age        SibSp   Parch
##  0:97     1:  0   female: 76   Min.   : 2.00   0:120   0:134
##  1:87     2:184   male  :108   1st Qu.:25.00   1: 55   1: 32
##           3:  0                Median :36.00   2:  8   2: 16
##                                Mean   :38.01   3:  1   3:  2
##                                3rd Qu.:48.00   4:  0   4:  0
##                                Max.   :86.00   5:  0   5:  0
##                                                8:  0   6:  0
##       Fare       Embarked
##  Min.   : 0.00   C: 17
##  1st Qu.:13.00   Q:  3
##  Median :14.25   S:164
##  Mean   :20.66
##  3rd Qu.:26.00
##  Max.   :73.50
##
```

```
boxplot(passenger2class$Fare)
```




```
plot(passenger2class$Age, passenger2class$Fare,  xlab="Edat", ylab="Fare")
```



Igual que en el cas anterior, observem que hi han persones amb edat inferior a 20 anys que han pagat un preu elevat.

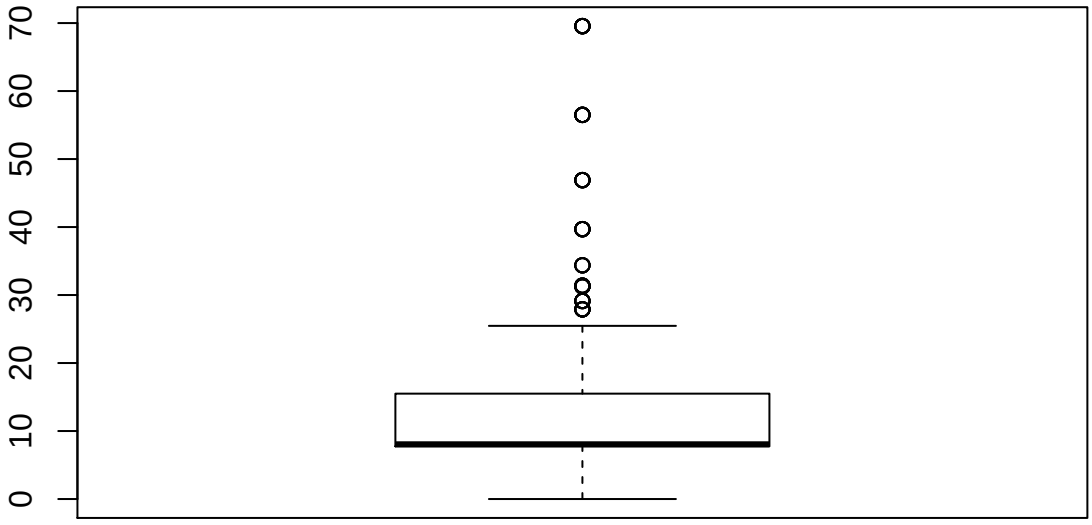
Fem el mateix estudi amb els passatgers de 3 classe.

```
passenger3class <- titanicdataset[titanicdataset$Pclass == "3",]
summary(passenger3class)
```

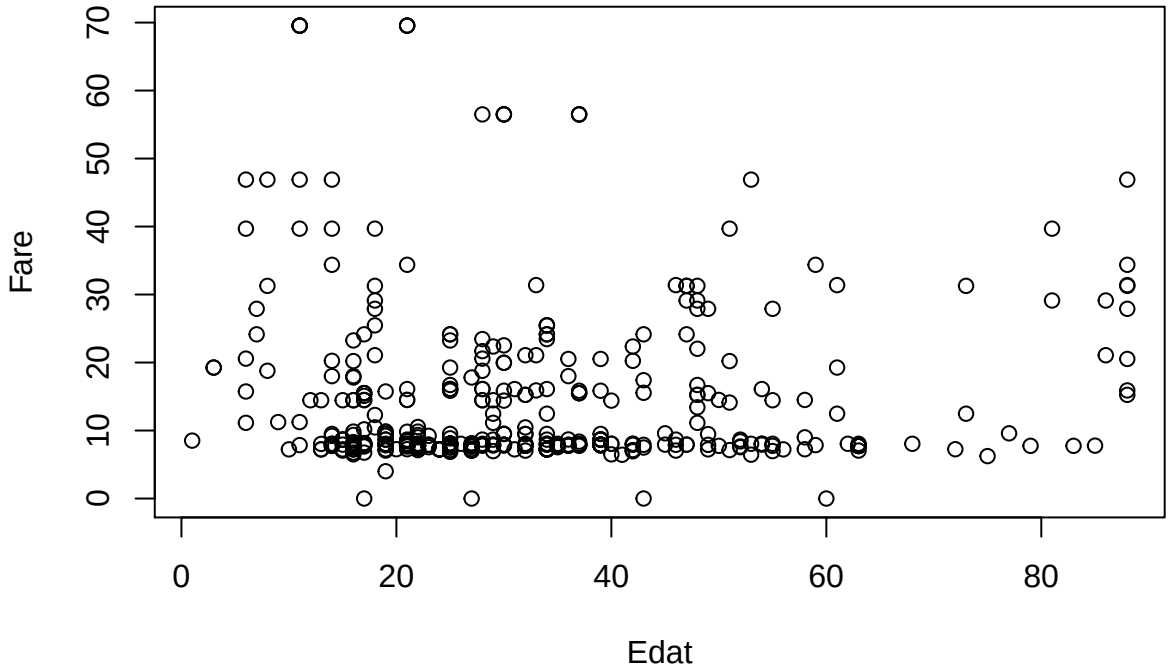
```
##  Survived Pclass      Sex           Age        SibSp   Parch
##  0:372    1:  0   female:144   Min.   : 1.00   0:351   0:381
##  1:119    2:  0   male  :347   1st Qu.:21.00   1: 83   1: 55
##           3:491                Median :28.00   2: 15   2: 43
##                                Mean   :32.19   3: 12   3:  3
##                                3rd Qu.:39.00   4: 18   4:  3
##                                Max.   :88.00   5:  5   5:  5
```

```
##                                                 8:  7   6:  1
##       Fare       Embarked
##  Min.   : 0.00   C: 66
##  1st Qu.: 7.75   Q: 72
##  Median : 8.05   S:353
##  Mean   :13.68
##  3rd Qu.:15.50
##  Max.   :69.55
##
```

```
boxplot(passenger3class$Fare)
```




```
plot(passenger3class$Age, passenger3class$Fare,  xlab="Edat", ylab="Fare")
```



Observem massa observacions amb valors atípics. Tenim molts passatgers de 3 classe que tenen un preu molt similar als passatgers de 2 classe. Definitivament podem pensar que aquest atribut te massa valors mal informats, i com no tenim manera de recuperar la informació original, descartarem el atribut.

```r
titanicdataset$Fare <- NULL
```

Amb aquest passos donem per finalitzada la fase de neteja. A continuació fem l'exportació del dataset amb les dades ja processades per utilitzar en les següents fases del projecte.

```r
write.csv(titanicdataset, "titanic_data_clean.csv", row.names = FALSE)
```