Antonio Nogueras i Alexandre Casanovas: M2.951 PRA 2 Els fets del Titanic

# 1.Descripció del data set: l'enfonsament del Titanic en el seu primer viatge

L’arxiu de partida és all_Titanic.csv.

Els camps són els següents:

**Name:** Nom del passatger o tripulant.

**Age:** Edat del passatger o tripulant.

**Boarded:** Port al que va embarcar.

**Position:** Per als tripulants és la seva ocupació, pels passatgers la seva classe (1classPass, 2classPass, 3classPass)

**Lifeboat:** Bot salvavides al que embarquen els supervivents, si el camp està vuit és que no van sobreviure.

**Body:** Indicador de que el cadàver del no supervivent es va recuperar.

**Sex:** Sexe del passatger o tripulant.

**Class:** Classe del passatger, en cas del tripulant classe 0.

**Group:** Grup al que pertany el passatger o tripulant.

El Titanic volia ser una demostració extrema de la tècnica i del luxe, que per això té tant de significat l'enfonsament del vaixell en el seu primer viatge amb passatgers. Portava tanta gent de l'alta societat, que el viatge era un fet social. Tant és així que alguns com Marconi van treure el bitllet d'aquest primer viatge encara que finalment la seva agenda no els va permetre abordar al Titanic.

Com al recent cas del Costa Concordia, pot ser que confiats per la seua superioritat tècnica, els oficials no van estar atents a les incidències. Per exemple,

- El Titanic quasi va col·lidir amb el transatlàntic City of New York a l'eixida de Southampton.
- Van rebre 6 avisos d’icebergs, que evidentment no es tingueren en compte: en el moment de la col·lisió, anaven a 22,5 nusos, sent el màxim del Titanic 23. Van veure l’iceberg i en 30 segons van col·lidir.

Hi havia molt de luxe, però no tant de disseny:

- Per què es van unir les planxes d'acer amb reblons? Els reblons probablement van cedir. L'iceberg no va trencar probablement les planxes, les va separar.
- És clar que no estava previst que el pes de sis compartiments inundats fóra suficient per a trencar-se en dos. Quan el pes de la proa va ser excessiu es va partir pel mig i enfonsar molt ràpidament.

Per altra banda, es pot pensar si l'extrema absència de mesures de seguretat va ser per criteris de la White Star Line (encara que dins de la reglamentació de 1912) : perquè portava a soles 20 bots, quan el disseny incloïa 64? 20 bots, dels quals 4 eren inferiors, no podien en cap cas amb 2208 persones que estaven a bord, i molt menys, amb la cabuda màxima del Titanic de 3547 persones.

Els bots, malgrat ser insuficients, no es van omplir: el primer d'ells portava 28 persones amb una cabuda de 65. Tardaren 1 hora a omplir el primer bot. Malgrat tot, del nostre anàlisi, quasi el 80% dels passatgers de primera classe, amb servei personal a bord també a 1a classe, es va salvar.

El dissenyador anava a bord, i havia de saber que amb sis compartiments inundats, el Titanic no tenia salvació: el màxim era 4. Actuaren com si, malgrat tot, no anés a afonar-se? Aprofitaran les dues hores, 40 minuts que va tardar a fer-ho?

Molts dels passatgers de les cobertes inferiors es quedaren atrapats amb portes estanques tancades per a contenir la inundació, la qual cosa va ser inútil perquè sis compartiments es van inundar i anava a afonar-se en qualsevol cas, i perquè es va trencar pel mig, i l'enfonsament va ser molt ràpid.

Dades que apareixen en https://titanicfacts.net/ Wikipedia o Encyclopedia Titanica.

*Creguem que hem exposat perquè és important. Quina pregunta pretenem respondre?*

Ja sabíem que no tots els grups es van salvar per igual. La idea és analitzar en les dades els fets del Titanic, els 'Titanic facts': fets com que, malgrat ser insuficients, els bots no anaven plens, especialment els més grans de 65 persones de cabuda. Fets com que la tripulació que no era marineria de coberta o oficials, es va ofegar quasi al 100% de mitja, al 100% en certs grups.

Fets com que els grups de més gran percentatge de salvats, els passatgers de primera classe que els acompanyava un servent, es va salvar el 78,6%, els mateixos servents d'ells el 70,7%, seguits de la tripulació de coberta amb 69,5%, molt per damunt de la 2a classe, de la 3a classe, de la resta de tripulació, dels quals 3 grups van morir tots.

## 2. Integració i selecció de les dades

Hi ha dades públiques de passatgers i tripulació del primer, i únic, viatge amb passatgers del Titanic, que hem triat analitzar.

Les dades dels passatgers i tripulació del Titanic es poden trobar en diversos llocs; nosaltres hem triat la Wikipedia. El motiu és que en altres llocs ens limiten a una part de les dades: per exemple en Kaggle de 891 passatgers, sense incloure la tripulació, i les dades estan ja transformades per a pràctiques de ML.

Altres, com l'enciclopèdia Titànica, no donen dades, per exemple, d'en quin bot ens va salvar o si el mort es va recuperar.

El més gran percentatge d'ofegats es troba entre la tripulació, l'aprovisionament, el restaurant de luxe, l'orquestra, l'Harland and Wolff Guarantee Group de les drassanes... grups que hem inclòs en el terme general de tripulació.

En Crew_of_the_Titanic la Wikipedia ens ofereix dades que ens permeten analitzar que va passar amb aquests grups tan castigats.

En Passengers_of_the_Titanic la Wikipedia ens ofereix dades de tots els passatgers, però també és on hem trobat dades completes que permeten definir subgrups, com els servents acompanyants dels passatgers de 1a classe viatjant també en 1a classe, el segon grup en % en salvar-se.

Aquestes dades apareixen en taules d'Excel integrades en el text web, en dos URL diferents:

https://en.wikipedia.org/wiki/Passengers_of_the_Titanic

https://en.wikipedia.org/wiki/Crew_of_the_Titanic

En la Wikipedia els passatgers apareixen en 3 taules, de 1a, 2a i 3a classe, i la tripulació, en diverses taules de divisió per grups generals, com Officer, Deck, Victualling, Enginnering o Restaurant.

Aquestes taules podrien haver-se obtingut via scraping, però en aquest cas és molt més fàcil copiar i pegar en Excel directament, ajuntar-los, llevar-les referències de la Wikipedia tipus [68] en el mateix Excel i llegir el fitxer Excel en R o Python per al nostre anàlisi.

Per comoditat, en el mateix Excel:

- hem prescindit de columnes que no aportarien molt a l'anàlisi, com el lloc de naixement de cada passatger o tripulant,
- hem unificat la posició de les columnes
- hem inclòs als passatgers en la columna *Position* (títol o càrrec de la tripulació), que ens permet incloure el treball dels servents acompanyants de la primera classe (els passatgers figuren com tal, passatgers de 1a, 2a i 3a classe com, per exemple, "*1classPass*")
- hem inclòs la tripulació en una nova classe, distinta de la 1a, 2a o 3a, *la classe* 0.
- Hem llevat en els noms les referències a la bibliografia de la Wikipedia, com [63], etc., on expliquen dades personals especials del passatger o tripulant.
- En *Age* trobaren uns pocs casos de nonats d'edat en mesos entre els passatgers, que passarem a una edat amb decimals en anys.
- La columna *Sex* s'ha generat comprovant en *Name,* mitjançant un *if,* si apareix la cadena 'Mr.' , o equivalents, com 'master', 'Colonel' per a sexe=m, i 'Mrs.' o 'Miss'... per al sexe=f. Per exemple, els xiquets són 'master', les xiquetes 'miss'.

# 3. Neteja de les dades

## 3.1 Zeros i buits

Hem provat amb dues aproximacions de neteja i de desenvolupament a causa del grau de detall en l'anàlisi: en la versió amb més columnes, el motiu de mantenir la versió estesa és perquè està més implicada en la demanda de "perquè és important el data set i quina pregunta pretén respondre": el necessitem per justificar l'elecció i les conclusions.

Però al mateix temps calia complir en totes les demandes de la PRA2, com la predicció i els models, per això presentem Titanic.rmd que resol la PRA2 en una peça més sintètica.

A la versió estesa, el fitxer s'havia fet massa llarg, perquè hi havia moltes coses que volíem ressaltar, la qual cosa és el motiu d'oferir fitxers separats per als punts 3, 4 i 5 de la PRA2 per tal de fer-ho més accessible.

La versió reduïda, Titanic.rmd, resol els tres punts de la PRA2 dins del RMD Titanic.

No hi ha zeros en les dades que no siguin valors numèrics vàlids:

- En Age fiquen NA en les absències de dada,
- Lifeboat i Body no inclouen zeros i els NA estan relacionats amb els ofegats, en Lifeboat, o ofegats que no s'ha trobat el cos, en Body.
- En Class hem ficat com a classe 0 la tripulació i hi ha 3 NA.

Els casos de NA són prou freqüents en dues columnes, Lifeboat i Body per les raons que hem anomenat i els hem conservat. En tot cas hem reduït, per no ser necessàries, el nombre de columnes en fer l'anàlisi i per exemple Body no es gasta.

Tasques que hem dut a terme:

- Els fitxers originals de la Wikipedia no inclouen directament els salvats, sinó que indiquen els bots en els quals se salvaren. Hem afegit una columna addicional Survived, 1 si s'ha salvat, 0 si s'ha ofegat. En Lifeboat hi ha 20 categories de bots que volem conservar perquè els bots no anaven plens.
- Hem estudiat duplicitats en els noms per si podien emprar-se com identificadors únics. Hem trobat que hi ha parelles de nom i cognoms en un total de 6 casos que, menys u, es tracta de nom i cognoms corrents que són diferents en l'edat, ofici, on han abordat el Titanic o en la classe, menys un cas del cuiner "Coutin, Mr. Auguste Louis, entree cook", de 29 anys que apareix alhora en Victualling i Restaurant, que hem eliminat u dels dos.
- Com el nom no identifica de manera única, hem afegit posteriorment una columna ID d'identificació numèrica.
- Hem comprovat els NA. Hi ha en u en Age, omplit amb la mitjana d'edat del seu grup, en Class i Boarded, corregit verificant en la Wikipedia . En la versió reduïda s'han llevat aquestes dades, i Boarded no se selecciona.
- A més, hi ha dues columnes, Lifeboat i Body, amb NA de forma natural, i no és necessari llevar els NA. En la versió reduïda no s'han seleccionat.
- Hem corregit els errors, i finalment hem comprovat el resultat abans de salvar el .csv allTitanic, en la versió llarga, en Titanic.csv en la reduïda.
- Hem creat una variable binària Adult, que pren els valors 0 o 1 segons siguin menors o majors de 15 anys.
- Hem creat variables acumulades cSurvived, cSex, cAdult, cClass0, cClass1, cClass2 i cClass3 per fer estudis de regressió lineal.

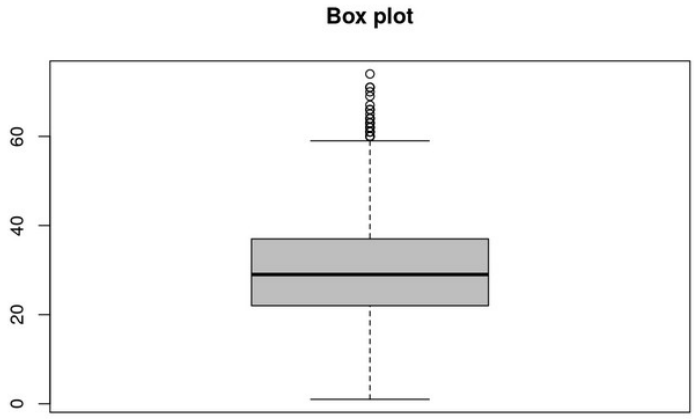
El procés queda reflectit, en la versió estesa, en el fitxer CleaningTitanic.ipynb que s'adjunta i en la versió reduïda al fitxer titanic.RMD.

### 3.2 Identificació i tractament de valors extrems.

Al fitxer de dades a soles pot haver-hi valors extrems en la única columna numèrica Age (l'identificador ID no afecta, la resta són factors).

En Age, el mínim és d'un nonat de 0,166 anys, i el màxim de 74, la qual cosa és raonable.

Mitjançant un gràfic tipus box plot hem fet un estudi visual.




La mitjana d'edat està a 30.09 i el mínim un nonat de 0,166 anys i el màxim en 74. El rang interquartil està entre 22 i 37 anys, amb el que el gràfic deixa com a outliers totes les edats per sobre de 60 anys. A la realitat sabem que això no és cert, no són edats fora d'el normal, simplement que la mitjana d'edat és baixa.

## 4. Anàlisi de les dades

### 4.1 Selecció dels grups de dades

Al fitxer .csv de dades hi ha dades informatives i dades útils per l'anàlisi, que seran les que seleccionem.

Dades informatives són si han abordat el Titanic a França, UK o Irlanda (Boarded) o el càrrec (Position). Dades d'interès poden ser l'anàlisi de l'edat (Age), els bots salvavides (Lifeboat), el sexe (Sex), el grup (Group), sobrevivents (Survived) o classe 1a, 2a o 3a en Class.

*En la versió reduïda* seleccionem com paràmetres Adult (0 si Age <15, 1 si Age >= 15), Class (0,1,2,3), Sex (m o f) i Survived (TRUE o FALSE). Per tant, les quatre són factor.

*En la versió estesa* es conserva Age (com única numèrica que ens permet tractar temes com la normalitat i homoscedasticitat), Lifeboat (20 bots salvavides: 1,2,...,16,A,B,C,D), Group (inclou els subgrups de la 1a classe així com els grups de tripulació, com Officer, Deck, Victualling, Enginnering o Restaurant), Class (0, la tripulació, 1, 2 o 3 segons la classe), Sex (m o f) i Survived (TRUE o FALSE).

Com a exemple de dades informatives, a Queenstown, Irlanda pujaren molts immigrants molts jóvens (23 a 25 anys), molts en família, en la tercera classe. Podem filtrar per Boarded=Q i Class="3" i analitzar els salvats per edat i sexe.

Dades d'interès per a l'anàlisi és analitzar el cas de la 1a classe que anaven acompanyats de secretaris, nurses, maids, valets, etc. que van tenir resultats diferents de la resta de 1a classe. Dades informatives serien, com veurem, que passatgers amb servents que van abordar el punt d'inici a Cherbourg, França (un poc a manera de creuer) el 94,1% se salvaren, i el 85% d'aquests servents que abordaren a Cherbourgs se salvaren així mateix.

Dades d'interès permeten comparar els resultats molt diversos de la tripulació, segons foren oficials o mariners de coberta amb relació a la resta de tripulació, com els de les màquines (Engineering) o personal de servei (Victualling).

En resum, el que interessa és la participació de cada u d'ells, o combinació de grups, en el reduït grup dels sobrevivents. Per exemple, la proporció de salvats menors de quinze anys, per cada u dels grups, entre els sobrevivents.

Per a realitzar les proves estadístiques hem creat 3 grups de dades segons l'estudi:

- Per un arbre de decisió, Survived, Sex, Adult i Class.
- Per una regressió lineal cSex, cSurvived, cAdult, cClass0, cClass1, cClasss2 i cClass3.
- Dividim la mostra en dues, Safe i Died, segons es van salvar o van morir i fem un test d'hipòtesis sobre la mitjana de l'edat.

## 4.2 Comprovació de la normalitat i homogeneïtat de la variància

Hem fet estudis de normalitat i homogeneïtat de la variància que es poden trobar al codi en R a l'apartat **«Comprovació de la normalitat i homogeneïtat de la variància».**

Els estudis sobre homogeneïtat de la variància s'han fet entre la variable Survived i les variables Adult, Class i Sex per una banda i entre la variable acumulada cSurvived i les variables acumulades cAdult, cSex, cClass0, cClass1, cClass2 i cClass3.

En tots els casos els p-values són tots molt propers a 0 i menors que 0.05, per tant podem rebutjar la hipòtesis nul-la, per tant podem dir que les variàncies no són iguals i assumim heterocedasticitat.

Les comprovacions sobre normalitat s'han fet sobre les variables acumulades cSurvived, cAdult, cSex, cClass0, cClass1, cClass2 i cClass3.

Els p-values són tots molt propers a 0 i menors que 0.05, per tant podem rebutjar la hipòtesis nul-la, per tant podem dir que les distribucions no són normals.

## 4.3 Aplicació de proves estadístiques.

Aquestes probes es poden trobar a titanic.RMD.

***Arbre de decisió*:**

```
tree_model <- C50::C5.0(trainX, trainy, rules=TRUE )
```

Un estudi per trobar les regles que van portar a sobreviure o no.

***Regressió lineal sobre les variables acumulades:***

```
model = lm(cSurvived ~ cSex, data = Titanic)

model = lm(cSurvived ~ cAdult, data = Titanic)

model = lm(cSurvived ~ cClass0, data = Titanic)

model = lm(cSurvived ~ cClass1, data = Titanic)

model = lm(cSurvived ~ cClass2, data = Titanic)

model = lm(cSurvived ~ cClass3, data = Titanic)

model = lm(cSurvived ~ cSex + cAdult + cClass0 + cClass1 + cClass2 + cClass3,
data = Titanic)
```

La construcció d'un model per poder predir mitjançant una regressió lineal la variable acumulada cSurvived.

Primer vam voler fer un estudi de regressió logarítmica sobre Survived, però vam veure que les variables tenien poca correlació i el model no era prou bo.

```
##                Adult       Class        Sex    Survived

## Adult     1.00000000 -0.23103560 -0.1520305 -0.07173064
## Class    -0.23103560  1.00000000  0.2511492 -0.01458246
## Sex      -0.15203049  0.25114922  1.0000000  0.45674303
## Survived -0.07173064 -0.01458246  0.4567430  1.00000000
```

Per això vam crear les variables acumulades que permetien fer comparacions del tipus, número d'homes per supervivents o número de persones de primera classe per supervivents. En aquest cas vam veure una forta correlació i vam crear un bon model.

```
##           cSurvived      cSex    cAdult   cClass0   cClass1   cClass2   cClass3
## cSurvived 1.0000000 0.9805499 0.9815914 0.7242037 0.6916556 0.8527876 0.9217838
## cSex      0.9805499 1.0000000 0.9427709 0.5906714 0.6966437 0.8878291 0.9403003
## cAdult    0.9815914 0.9427709 1.0000000 0.8236857 0.5636606 0.7619894 0.9389442
## cClass0   0.7242037 0.5906714 0.8236857 1.0000000 0.2328246 0.3675143 0.6535809
## cClass1   0.6916556 0.6966437 0.5636606 0.2328246 1.0000000 0.7304298 0.4534333
## cClass2   0.8527876 0.8878291 0.7619894 0.3675143 0.7304298 1.0000000 0.7119293
## cClass3   0.9217838 0.9403003 0.9389442 0.6535809 0.4534333 0.7119293 1.0000000
```

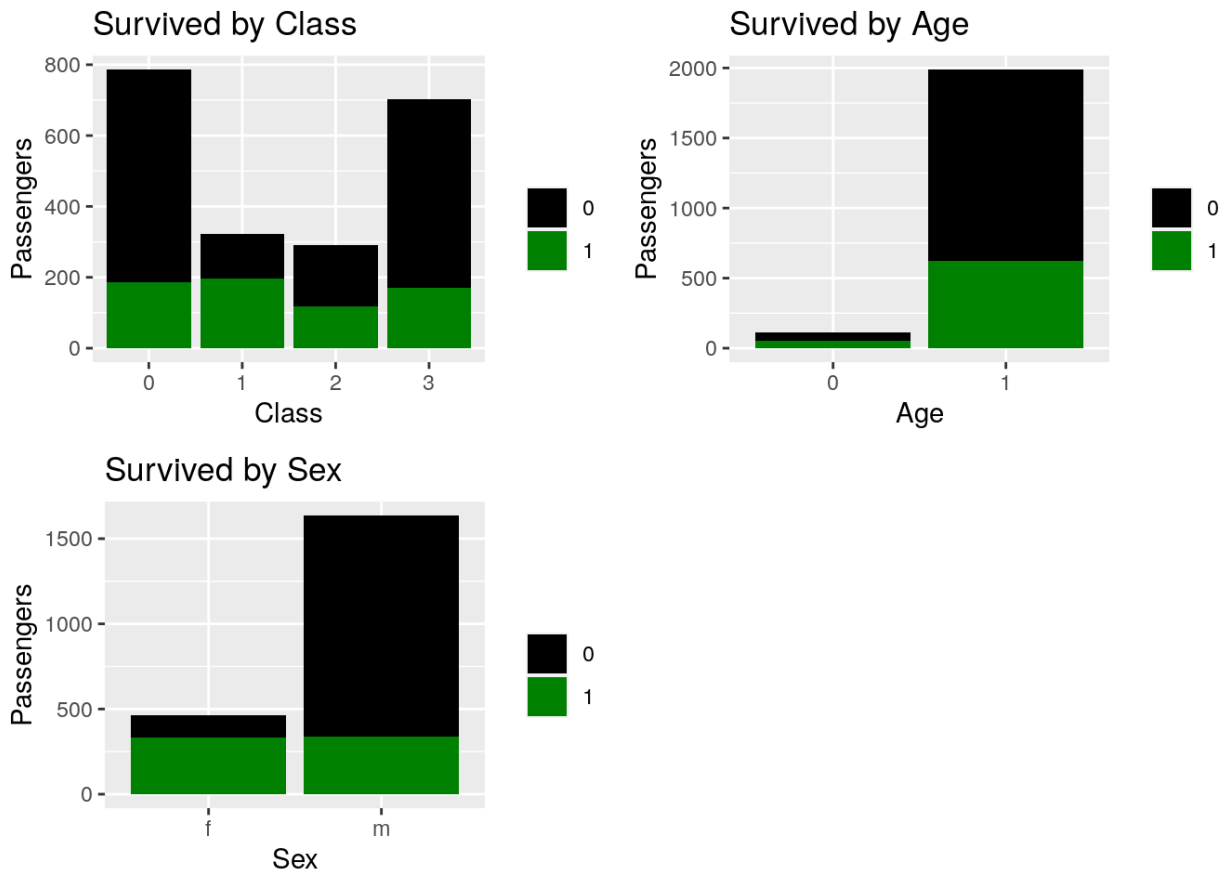
### Test d'hipòtesis de dues mostres independents sobre la mitjana:

```
t.test(Safe$Age,Died$Age,alternative="greater", var.equal=FALSE)
```

Un test d'hipòtesis per poder compara la mitjana d'edat entre els supervivents i els morts.

## 5. Gràfiques

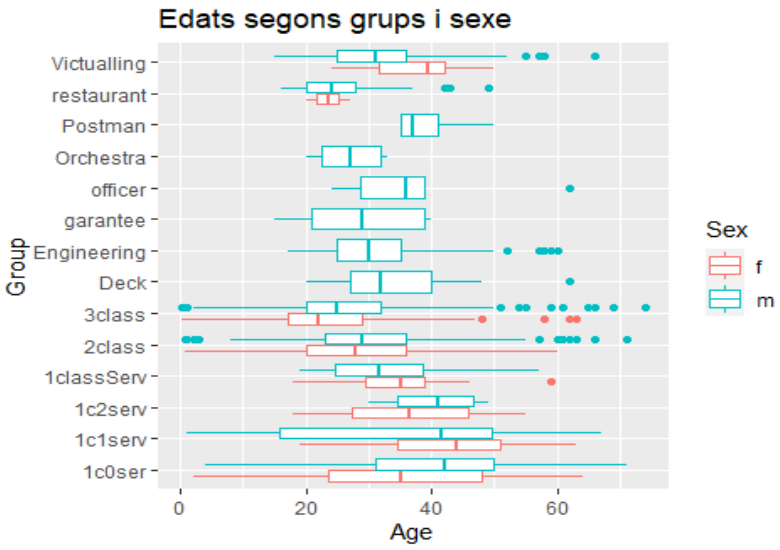
*En la versió reduïda:*

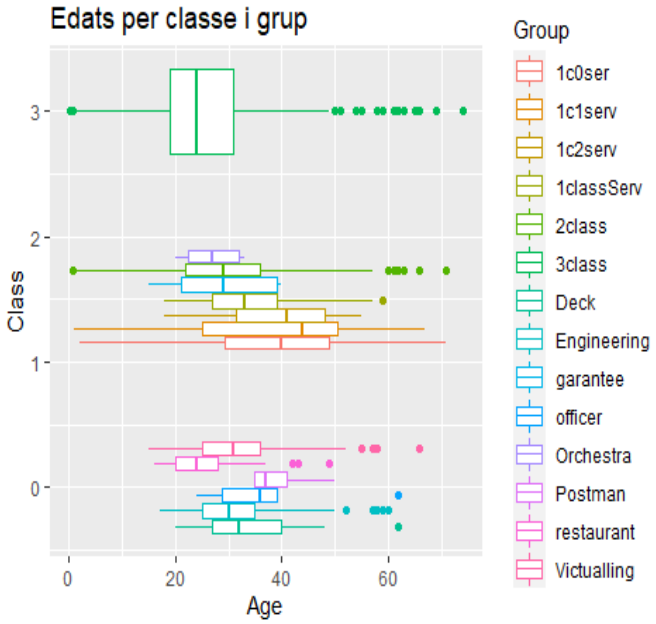


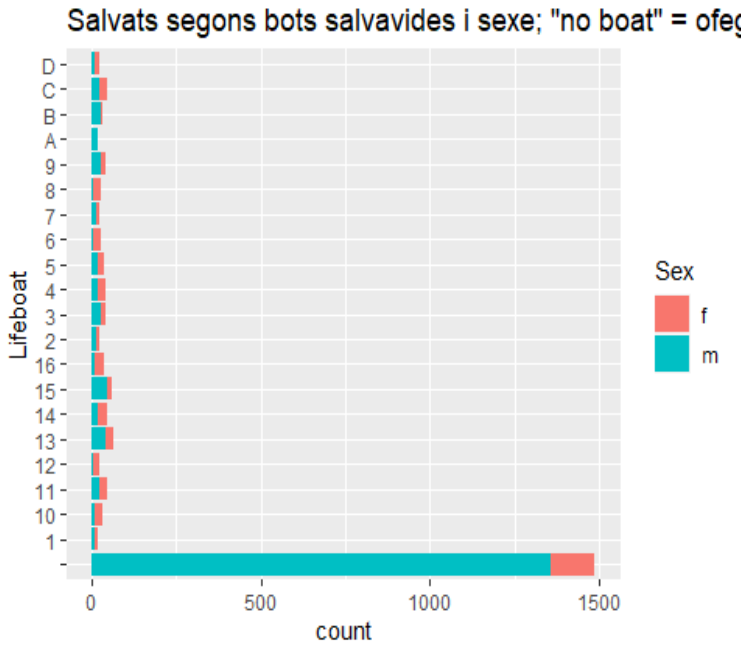

*En la versió estesa:*

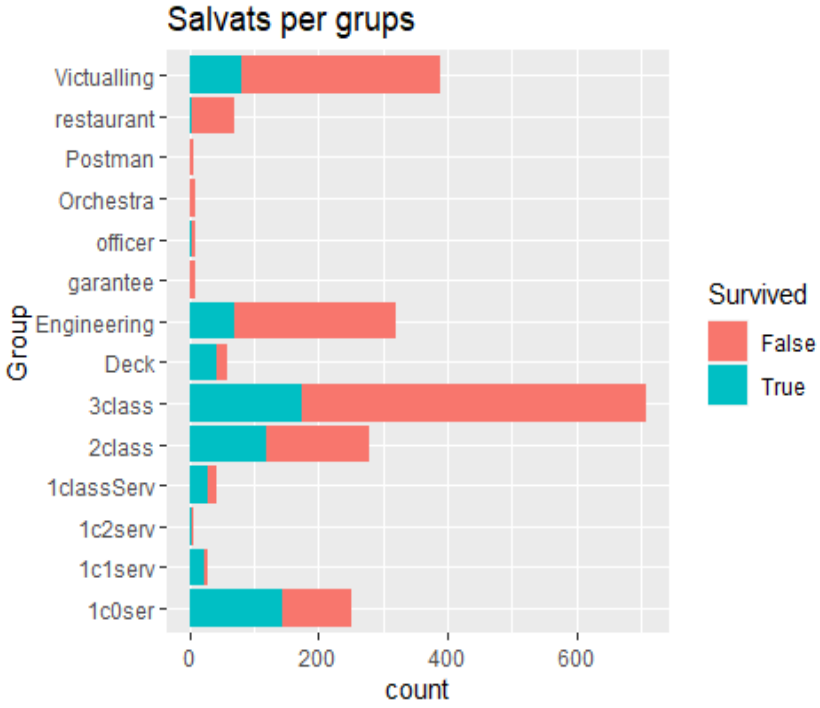
Els resultats complets d'aquest apartat estan recollits als fitxers 'Titanic_gràfiques.Rmd' i 'Titanic_gràfiques.html'

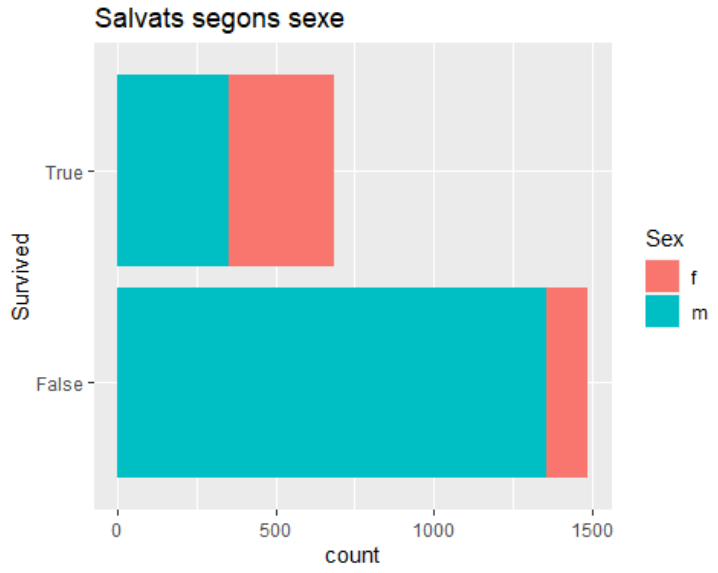
Algunes de les gràfiques d'aquest fitxer:

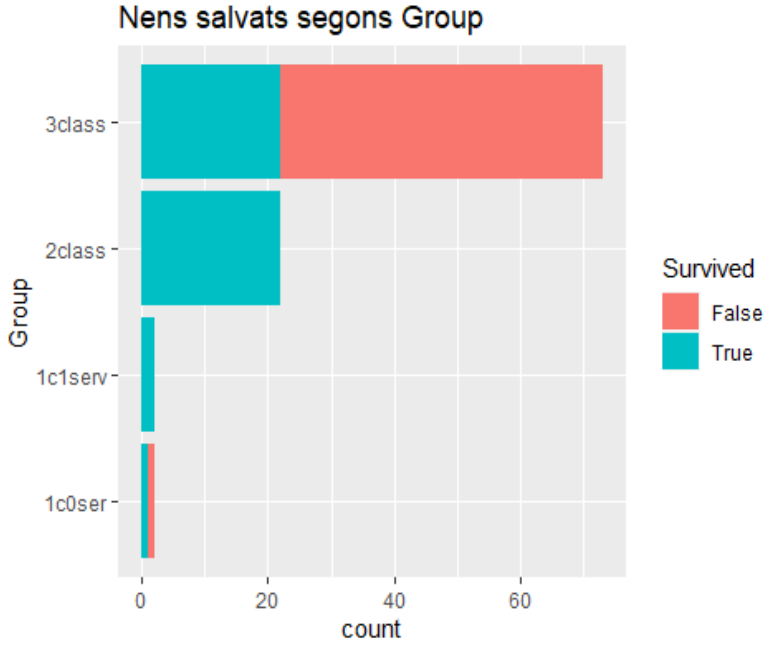

Edats segons grups i sexe
Victualling
restaurant
Postman
Orchestra
officer
garantee
Engineering
Deck
3class
2class
1classServ
1c2serv
1c1serv
1c0ser
Group
0
20
40
60
Age
Sex
f
m

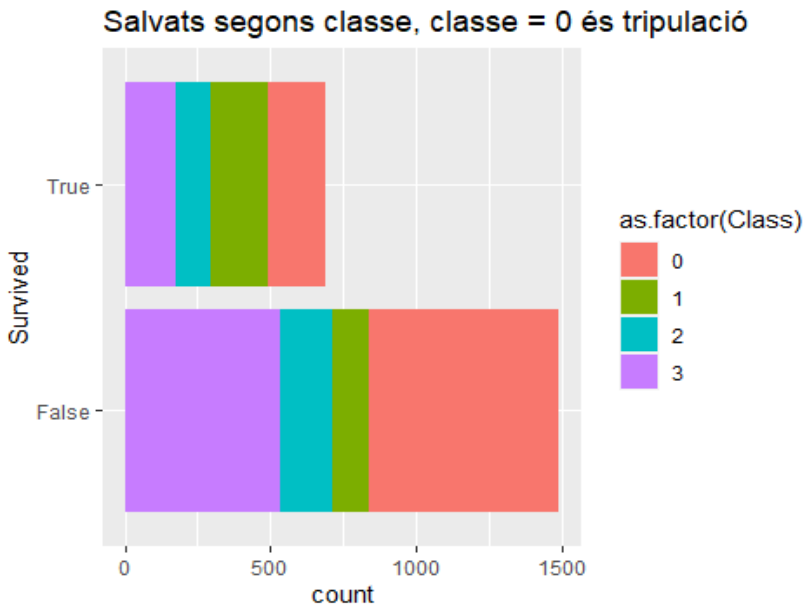

Edats per classe i grup
3
2
1
0
Class
0
20
40
60
Age
Group
1c0ser
1c1serv
1c2serv
1classServ
2class
3class
Deck
Engineering
garantee
officer
Orchestra
Postman
restaurant
Victualling

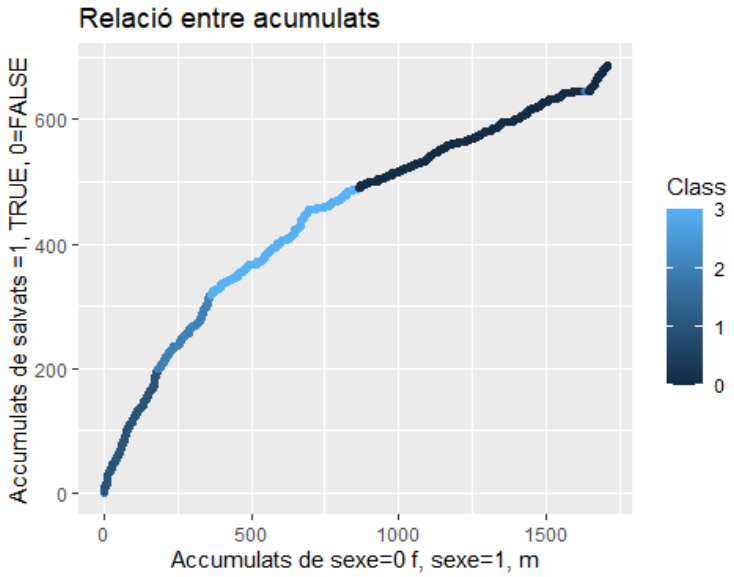

Salvats segons bots salvavides i sexe; "no boat" = ofe
Lifeboat
D
C
B
A
9
8
7
6
5
4
3
2
16
15
14
13
12
11
10
1
-
0
500
1000
1500
count
Sex
f
m


Salvats per grups
Group
Victualling
restaurant
Postman
Orchestra
officer
garantee
Engineering
Deck
3class
2class
1classServ
1c2serv
1c1serv
1c0ser
0
200
400
600
count
Survived
False
True


Salvats segons sexe
Survived
True
False
0
500
1000
1500
count
Sex
f
m


Nens salvats segons Group
Group
3class
2class
1c1serv
1c0ser
0
20
40
60
count
Survived
False
True



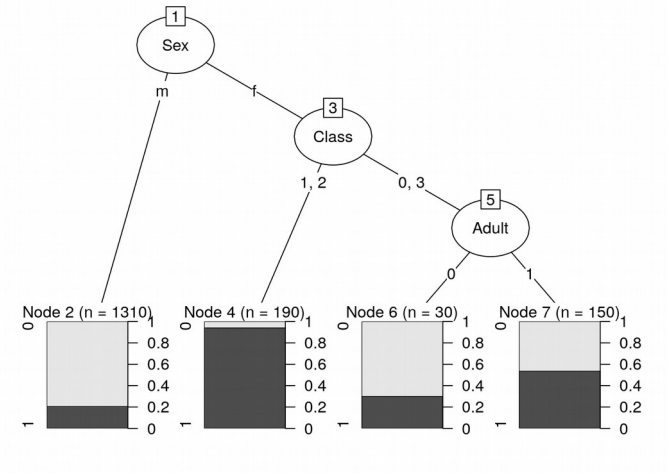




El pendent que s'obté és aproximadament la fracció del total d'homes que se salva.

# 6. Resolució de la pregunta: els fets del Titanic, els 'Titanic facts'.

Dins de les 3 proves que hem fet volíem respondre 3 preguntes:

**Quines són les regles principals que porten a sobreviure o no?**




L'arbre de decisió classifica malament 366 casos un 21,8%.

Poden extreure les següents regles:

- Els homes moren en un 78,8%
- La tripulació i la classe tercera moren en un 75%.
- Les dones de primera i segona classe es salven en un 93,8%
- Les dones adultes es salven en 75,8%

Hem vist que no hi ha cap regla que digui que els nens es salvin més que els adults, al contrari amb les dones es salven més les adultes.

Això ens ha portat a la següent pregunta.

**La mitjana d’edat dels supervivents és igual a la dels no supervivents?**

Per a comprovar-ho hem fet un tes d'hipòtesis de dues mostres independents sobre la mitjana d’edat dels que es van salvar i els que no.

Partim de les hipòtesis nul-la i alternativa següents.

H0: Els no supervivents tenen igual edat que els supervivents.

H1: Els no supervivents tenen més edat que els supervivents.

El resultat ens dona un p-value de 0.7078 molt, més gran que 0,05 i no podem rebutjar la hipòtesis nul-la.

Per tant veiem que la mitjana d’edat dels supervivents és similar a la dels no supervivents.

**Podem crear un model prou bo per predir qui sobreviu i qui no?**

En aquest cas hem creat el següent model de regressió lineal:

```
model = lm(cSurvived ~ cSex + cAdult + cClass0 + cClass1 + cClass2 + cClass3,
data = Titanic)
```

Te un coeficient de determinació de 0.996, per tant és un model prou bo.

Per tant podem predir linealment el nombre de supervivents, mitjançant la resta de variables acumulades.