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

1.1 Introducció: el tema, i perquè és important

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.

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. A més, 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 i l'iceberg probablement va separar les planxes. 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 es va 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. 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. Actuaren com si, malgrat tot, no anés a afonar-se?

Molts quedaren atrapats en les cobertes inferiors per les portes estanques tancades per a contenir la inundació, la qual cosa va ser inútil, perquè anava a afonar-se en qualsevol cas pels sis compartiments inundats. En trencar-se pel mig, l'enfonsament va ser molt ràpid.

Hi ha un cert paral·lelisme en totes les tragèdies, i volem analitzar les dades d'aquesta.

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

1.2 Quina pregunta pretenem respondre?

La idea és analitzar en les dades els fets del Titanic, els 'Titanic facts': fets com les brutals diferencies en el percentatge de salvats segons els grups, o el fet que, malgrat ser insuficients, els bots no anaven plens, especialment els més grans de 65 persones de cabuda.

2. Selecció i integració de les dades

2.1 Selecció

Del primer, i únic, viatge amb passatgers del Titanic, hi ha dades públiques de passatgers i tripulació que 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 a dades de 891 passatgers i sense incloure la tripulació, i les dades estan ja transformades per a pràctiques de ML.

En *Crew_of_the_Titanic*, https://en.wikipedia.org/wiki/Crew_of_the_Titanic, la Wikipedia ens ofereix dades que ens permeten analitzar que va passar amb la tripulació.

Passengers_of_the_Titanic, https://en.wikipedia.org/wiki/Passengers_of_the_Titanic, la Wikipedia ens ofereix dades completes de tots els passatgers, que permeten definir subgrups de la 1a classe, que no era uniforme.

2.2 Integració de les dades

Hem ajuntat 10 fitxers parcials de la Wikipedia en un de 2178 dades de les següents variables:

Name: nom del passatger o tripulant (ch)

Age: edat del passatger/tripulant (dbl)

Boarded: port on van abordar el vaixell (B, C, S, Q) (ch)

Position: càrrec del tripulant/ servent/classe (ch)

Lifeboat: bot salvavides on es van salvar (1, 2,..16, A, B, C, D) (ch)

Body: cos de l'ofegat, si es va trobar (ch)

Sex: sexe del passatger o tripulant (m, f) (ch)

Class: classe del passatger (1, 2, 3); zero si és tripulant (dbl)

Group: grup de la tripulació/subgrup del passatge (ch)

Aquestes dades apareixen en la Wikipedia en taules d'Excel integrades en el text web, en dos URL diferents: 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 les taules Excel directament.

La integració de les dades es fa en el mateix Excel i fitxer Excel serà llegit pel R o Python en la nostra anàlisi.

3. Neteja de les dades

3.1 Zeros, buits i canvis de columna

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, per poder ajuntar-les.
- Per evitar NA futurs, hem inclòs als passatgers en la columna *Position* (títol o càrrec de la tripulació) Els passatgers de 1a, 2a i 3a classe figuren com tal, com per exemple, "*1classPass*")
- Per la mateixa raó, hem inclòs la tripulació en una 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 entre els passatgers uns pocs casos de nonats d'edat en mesos, que passarem a una edat en anys amb decimals.
- 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'.

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.
- Els NA de Body esta relacionat amb els cosos trobats, i no hem fet us de aquesta columna.
- En Class les zeros són la classe de la tripulació i hi ha NA per absència de dades.
- Els casos de NA són prou freqüents en Lifeboat i Body. El Lifeboat 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, que en tot cas, no és usada.
- Hi ha dues versions una, 'estesa' inclou Age, Boarded i Lifeboat, i la 'reduida' que prescindeix d'aquestes columnes.
- Hem comprovat els NA en Age, omplits amb la mediana d'edat del seu grup, en Class i Boarded, corregit verificant en la Wikipedia en els dos casos.
- Hem corregit els errors, i hem salvat el fitxer .csv: allTitanic.csv, en la versió llarga, Titanic.csv en la reduïda (el nombre de columnes és diferent).

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. Dades de t.test() i boxplot.stats() d'Age:

```
##  One Sample t-test
##
## data:  allT$Age
## t = 118.16, df = 2175, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  29.45391 30.44805
## sample estimates:
## mean of x
##  29.95098
```

- Llistat dels que estan fora ($out):

```
boxplot.stats(selT$Age)

## $stats
## [1]  1 22 29 36 57
## $n
## [1] 2176
## $conf
## [1] 28.52581 29.47419
## $out
##  [1] 59.000 62.000 71.000 60.000 61.000 59.000 64.000 59.000 70.000 64.000
## [11] 64.000 60.000 60.000 71.000 59.000 58.000 63.000 58.000 64.000 64.000
## [21] 59.000 61.000 62.000 61.000 61.000 63.000 60.000 59.000 58.000 62.000
```

```
## [31]  0.916 59.000 58.000 62.000 67.000 63.000 60.000  0.833 62.000 63.000
## [41] 60.000 61.000 71.000 63.000 60.000  0.750  0.833 66.000  0.833  0.750
## [51] 61.000 59.000  0.250  0.166 65.000 63.000 61.000  0.580 62.000 69.000
## [61] 58.000 74.000  0.420 63.000 66.000 62.000 58.000 59.000 60.000 58.000
## [71] 66.000 62.000
```

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).

En resum, el que interessa és la participació de cada u dels grups, o combinació de grups, en el grup dels sobrevivents. Per exemple, la proporció de salvats menors de quinze anys, per cada u dels grups, entre els sobrevivents, o si podem predir si se salven, o no, d'acord amb les dades.

4.2 Comprovació de la normalitat i homogeneïtat de la variància

Moltes proves estadístiques suposen distribució gaussiana de dades, o l'aproximació d'un nombre molt elevat de dades que aproxima les distribucions a la gaussiana. Alguns resultats d'aquest apartat depenen del reduït nombre de casos disponibles per a certes categories.

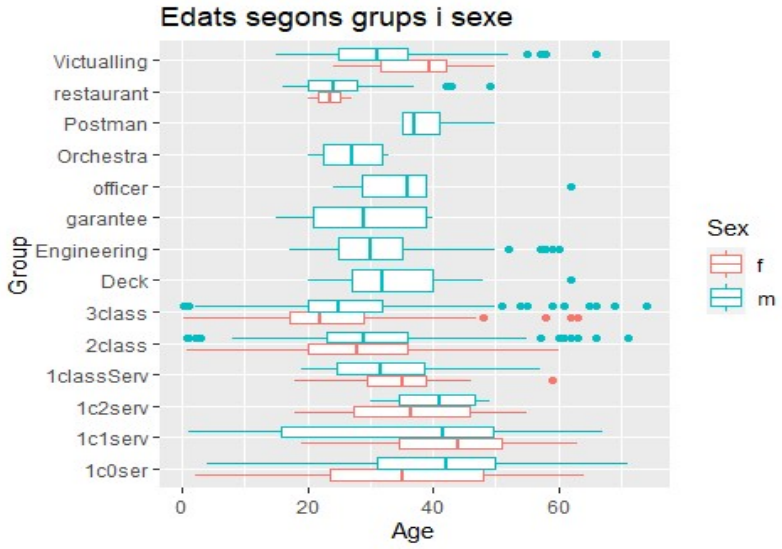
L'única columna que és numèrica, no categòrica, és l'edat, "Age"[1]. La distribució no pot ser gaussiana, ja que els valors menors que zero no poden existir, i els més grans, evidentment no arriben a infinit. En tot

[1] En el Titanic train.csv de Kaggle, amb dades de 891 passatgers, el problema és semblant: els grups són reduïts, i per tant les mostres, i les úniques columnes que són numèriques, no categòriques, és l'edat, "Age", i el preu del bitllet, "Fare", que és un categòric amagat, perquè els preus corresponen a les classes o nivells en la classe, com la coberta o cabina. Les dades, en el fons, són les mateixes, només la mostra és més reduïda i limitada a passatgers.

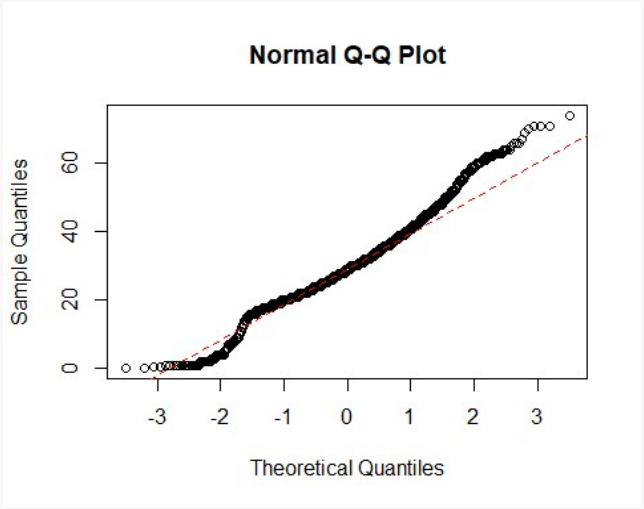
cas la mitjana és 29,95 anys i la desviació típica d'11,83, amb la qual cosa són marginals (outliers) d'edat per a una gaussiana els majors de 65,42 anys, dels que trobem 10 casos:

	ID	Name	Age	Boarded	Position
47	48	Artagaveytia, Mr. Ramon	71.0	C	1classPass
101	102	Crosby, Captain Edward	70.0	S	1classPass
144	145	Goldschmidt, Mr. George B.	71.0	C	1classPass
315	316	Straus, Mr. Isidor	67.0	S	1classPass
509	510	Mitchell, Mr. Henry Michael	71.0	S	2classPass
595	596	Wheadon, Mr. Edward H.	66.0	S	2classPass
1151	1152	Risien, Mr. Samuel Beard	69.0	S	3classPass
1239	1240	Svensson, Mr. Johan	74.0	S	3classPass
1288	1289	Webber, Mr. James	66.0	S	3classPass
2018	2019	Willis, Mr. William	66.0	S	Third Class Steward

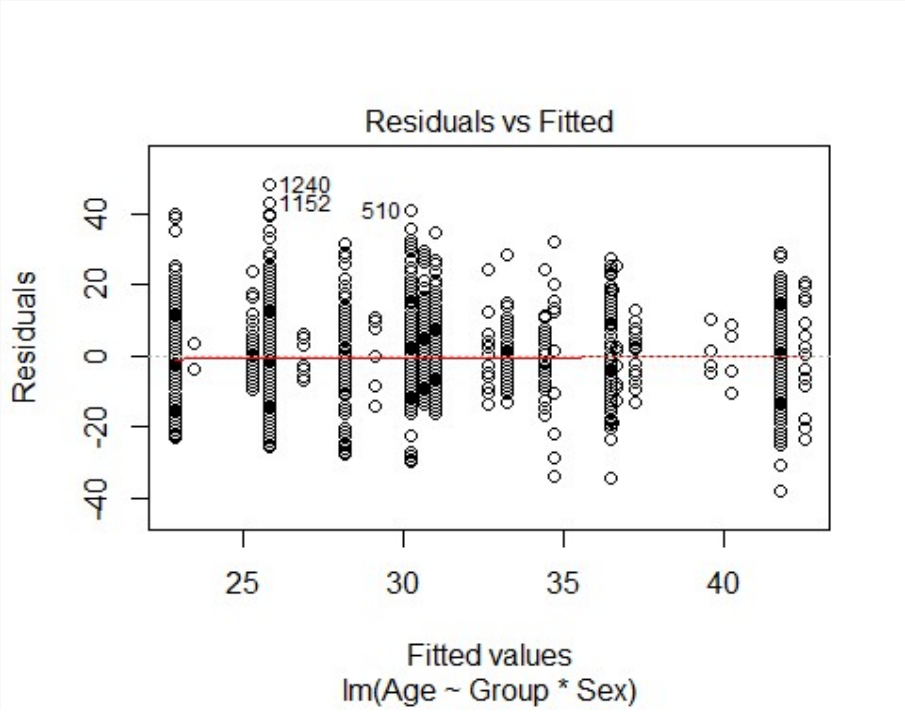
Els outliers d'Age per box-plot:



El qqplot amb la proximitat de la linea amb les dades ens diu que la normalitat es pot assumir menys a les cues:



La normalitat de la distribució dels residus i homoscedasticitat:



En la versió reduïda:

Hem fet estudis d'homogeneïtat de la variància entre la variable Survived i les variables Adult, Class i Sex i entre la variable acumulada cSurvived i les variables acumulades cAdult, cSex, cClass0, cClass1, cClass2 i cClass3. En tots els casos els valors-p són tots menors que 0.05 i podem dir que les variàncies no són iguals i assumim heteroscedasticitat.

Hem fet estudis de normalitat sobre les variables acumulades cSurvived, cAdult, cSex, cClass0, cClass1, cClass2 i cClass3. Els valors-p són tots menors que 0.05 i podem dir que les distribucions no són normals.

4.3 Aplicació de proves estadístiques

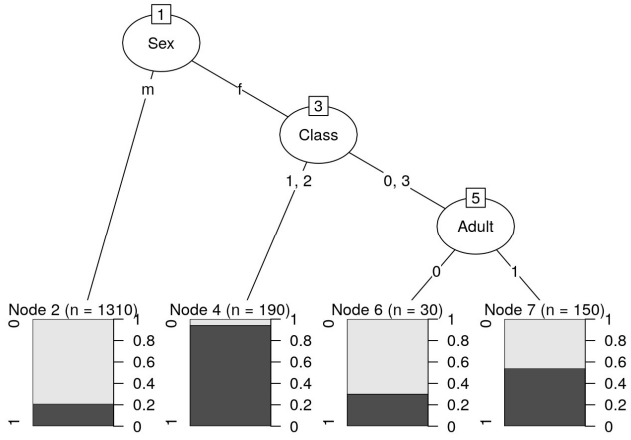
Fem ací un llistat de proves estadístiques, que estan adequadament explicades en els rmd i html.

Probes estadístiques en la versió reduïda, titanic.RMD:

Creació de models d'arbre, de regressió i t-test de diferència de mitjanes d'edat.

1-*Arbre de decisió,* ***per trobar les regles que van portar a sobreviure, o no:***

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



2-*Predicció amb el model d'arbre:*

```
predicted_model <- predict(tree_model, testX, type="class")
## [1] "La precisión de l'arbre és: 79.8100 %"
```

3-***Correlació de variables:*** Per fer un estudi de regressió logarítmica sobre Survived, però la correlació és baixa en les no acumulades:

```
##                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
```

Però una forta correlació en les variables acumulades, 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
```

4-*Regressió lineal sobre les variables acumulades, per poder predir mitjançant una regressió lineal:*

Un estudi amb lm(), entre cSurvived ~ cSex, cSurvived ~ cAdult, cSurvived ~ cClass0, cSurvived ~ cClass1, cSurvived ~ cClass2, cSurvived ~ cClass3 i lm(cSurvived ~ cSex + cAdult + cClass0 + cClass1 + cClass2 + cClass3, data = Titanic)

5-*Test d'hipòtesis t.test() de dues mostres independents sobre la mitjana, comparant la mitjana d'edat entre els supervivents i ofegats:*

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

El t.test rebutjava que hi haguera diferència de les mitjanes d'edat entre salvats i ofegats.

6-***Var.test (versió reduïda)*** comparant Survived contra les altres, (Class, Sex, Adult), 3 tests, com:

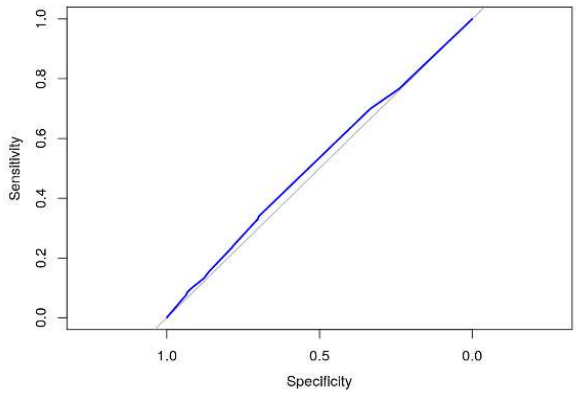
var.test(as.numeric(Titanic_S$Survived), as.numeric(Titanic_S$Adult))

F test to compare two variances

7-Corba de característiques operatives del receptor ROC

Corba ROC, que reflecteix la capacitat de diagnòstic d'un sistema classificador binari en variar el seu llindar de discriminació:

```
library(pROC)
roc <- roc(Titanic_S_R$Survived,reg_log_model$fitted.values)
## Area under the curve: 0.521
```



Probes addicionals en la versió estesa, allTitanic.rmd:

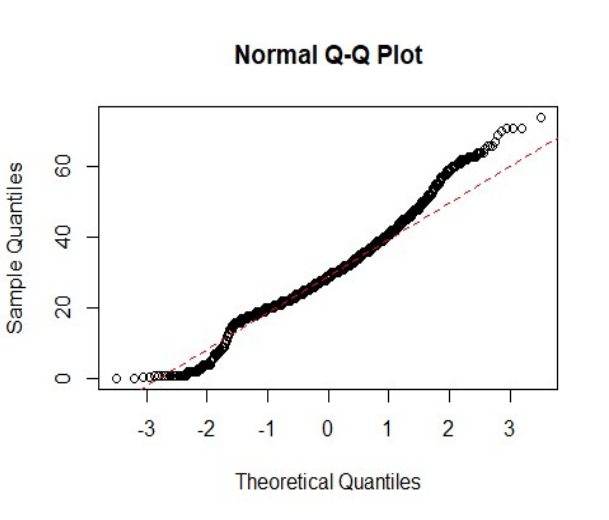
Amb la variable numèrica Age:

```
1-t.test(selT$Age)
```

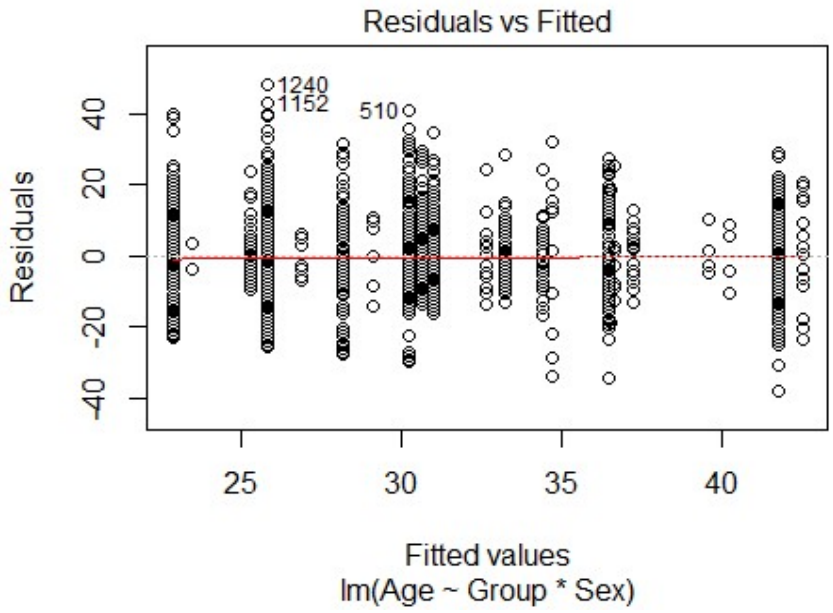
```
##  One Sample t-test
```

```
2-boxplot.stats(selT$Age)
```

3-Normal qqplot() d'Age:



4- Normalitat de la distribució dels residus i homoscedasticitat:



5- t-test comparant mitjanes de grups:

```
##  Welch Two Sample
```

6-*Kruskall Wallis* comparant mitjanes d'edat de la 1a classe i la 3a classe:

```
##  Kruskal-Wallis rank sum test
```

Les següents proves hem aplicat la tècnica que expliquem: per comparar les variables categòriques emprem la tècnica de cumsum(). Considerem variables categòriques amb dos nivells, 0 i 1; per exemple Survived (1=TRUE) i Class (1 i 2). La suma acumulada cumsum() dóna el nombre de salvats respecte a les altres variables acumulades, aconseguint variables numèriques en un gran nombre de valors diferents d'un llistat de variables factor.

7-Regressió amb lm()

```
## lm(formula = c1$cSurvived ~ c2$cSurvived)
##
## Multiple R-squared:  0.9861, Adjusted R-squared:  0.9861
## F-statistic: 2.092e+04 on 1 and 294 DF,  p-value: < 2.2e-16
```

El pendent dóna la relació entre el percentatge de salvats de 1a classe i 2a classe

8-*Constrast no paramètric de Survived de les classes 1a i 2a:wilcox.test*

```
##  Wilcoxon rank sum test with continuity correction
```

9-*Model de regressió lineal de salvats respecte a la 1a classe i predicció*:

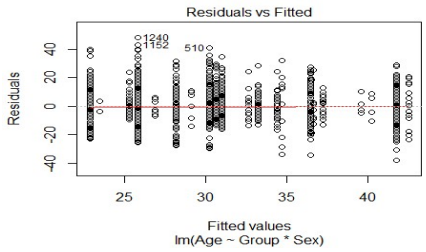
```
## lm(formula = cSurvived1 ~ cClass1, data = acumC12)
## Multiple R-squared:  0.9977, Adjusted R-squared:  0.9977
## F-statistic: 1.277e+05 on 1 and 294 DF,  p-value: < 2.2e-16
```

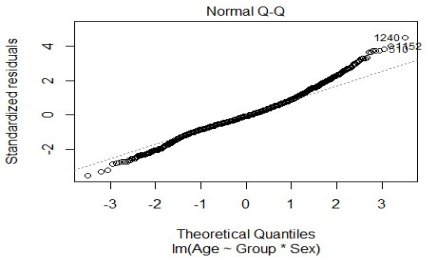
10-*Predicció de salvats de 2a classe en set casos aleatoris*:

```
## [1] "Predicted == valors de el fitxer test?" 7 casos
##  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE
```

11-*La homoscedasticitat dels residus per gràfiques "Residuales vs Fitted"*

Normalitat de la distribució dels residus i homoscedasticitat





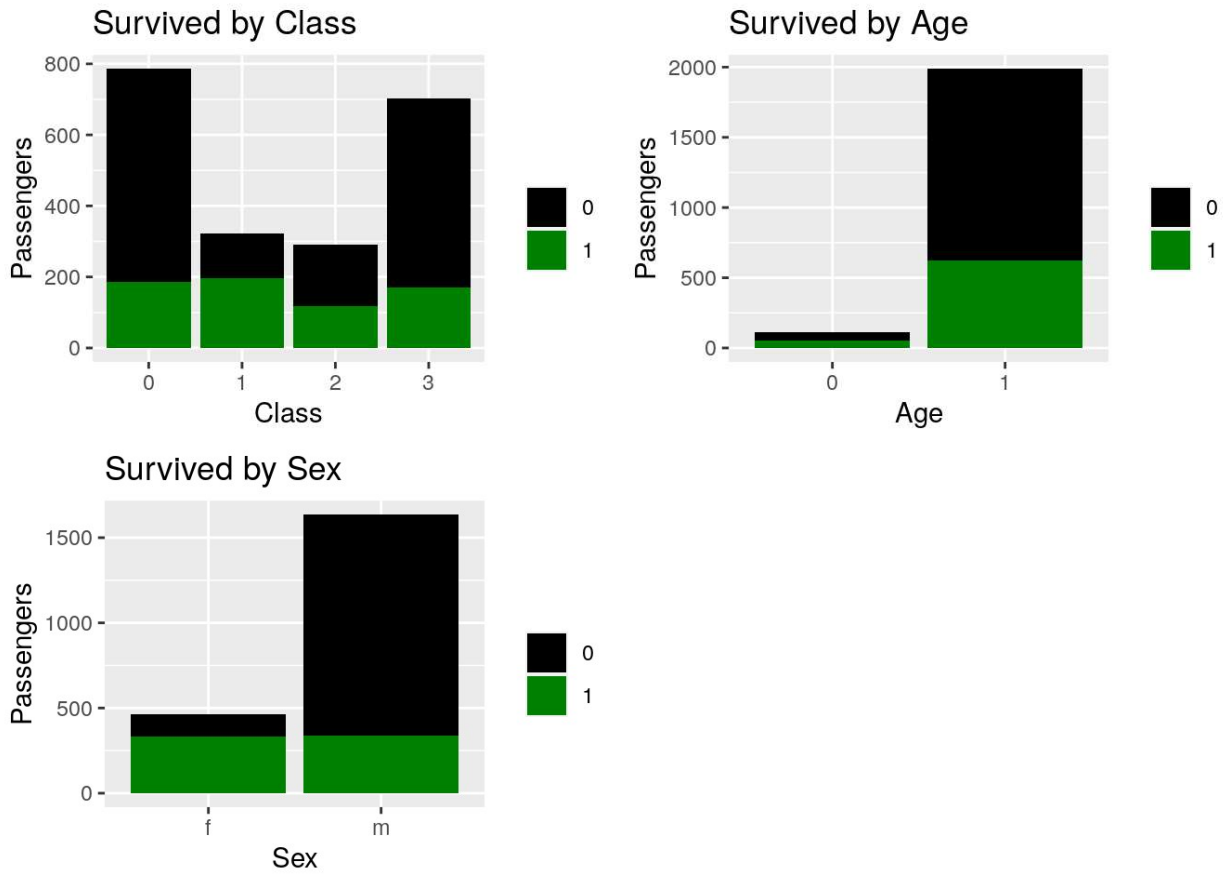
12- *Comprovació si el resultat es veu afectat per interacció de grups, com Group i Sex*

```
## lm(formula = Age ~ Group * Sex, data = allT)
##
## Multiple R-squared:  0.1764, Adjusted R-squared:  0.1684
## F-statistic: 21.97 on 21 and 2154 DF,  p-value: < 2.2e-16
```

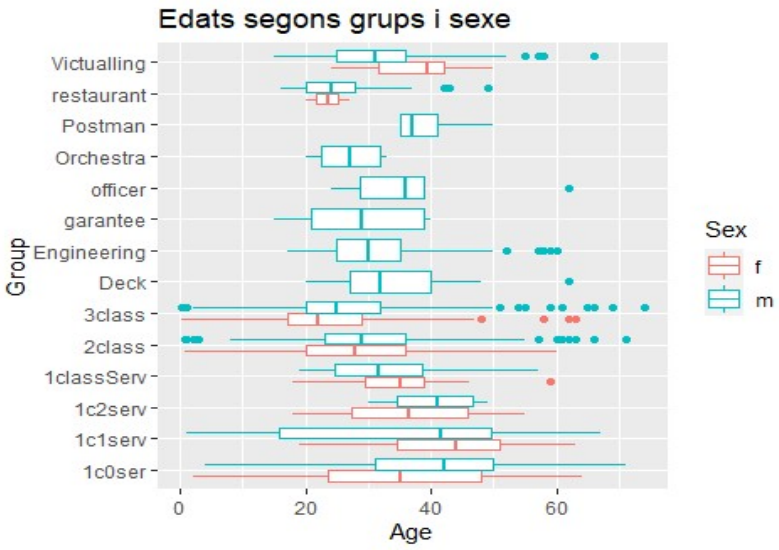
5. Gràfiques i taules

5.1 Gràfiques

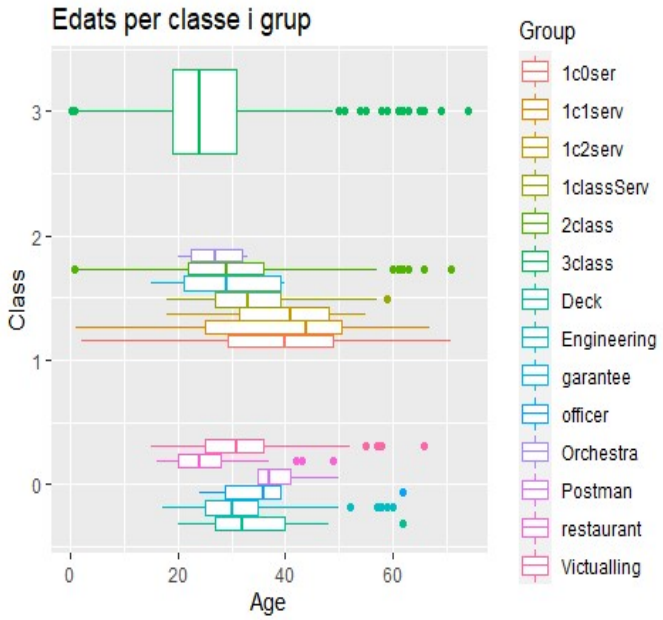
En la versió reduïda, bar plots de salvats per classe, edat i sexe:



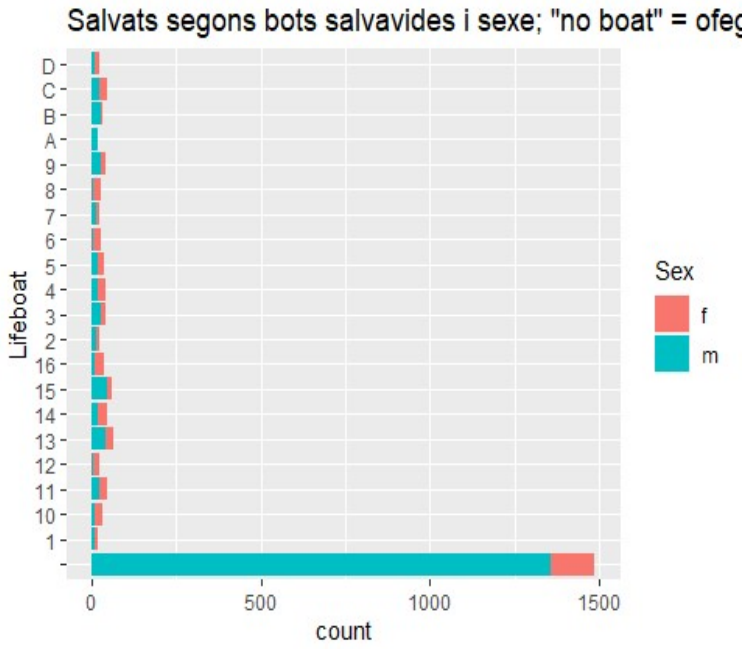
Algunes de les gràfiques de la versió estesa:



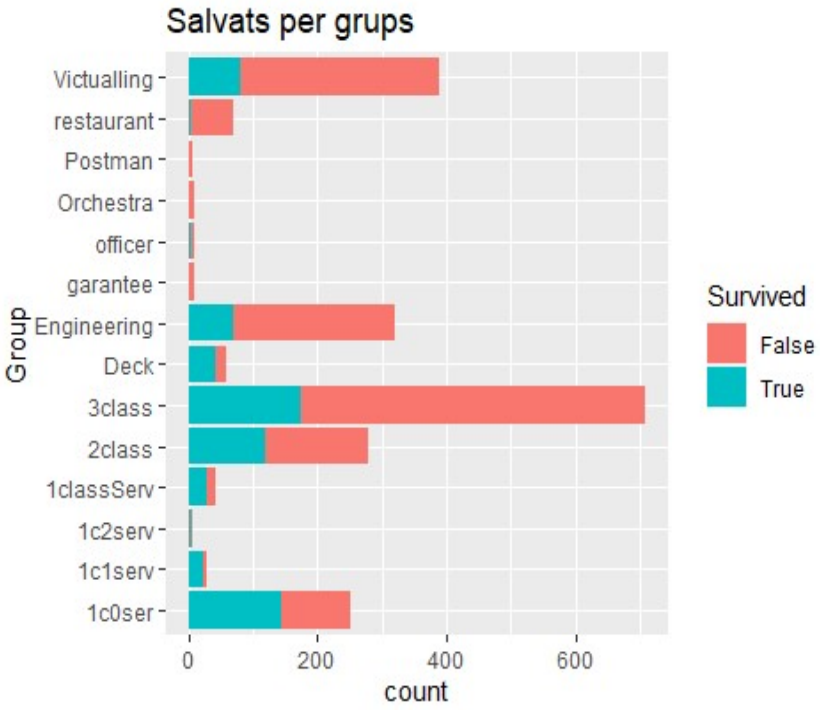
1- Ouliers d'edat segons grup i sexe



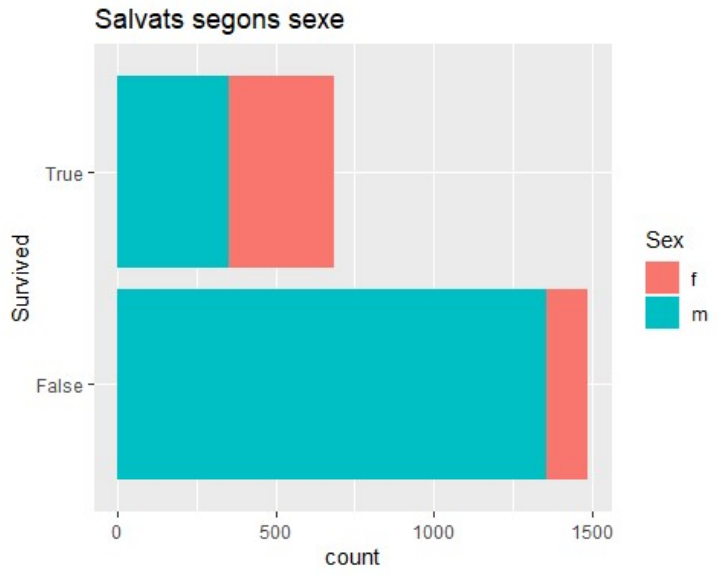
2- Ouliers d'edat segons grup i classe



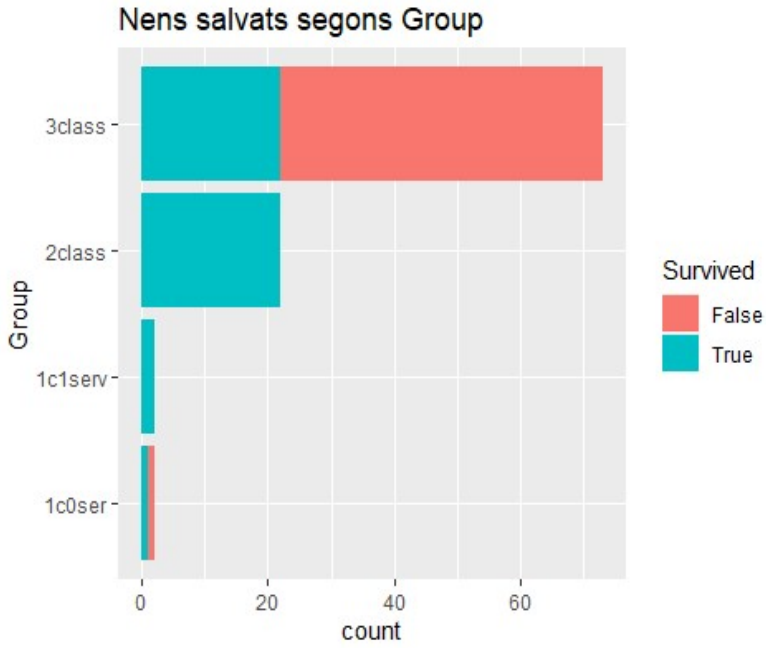
3- Salvats I ofegats segons sexe



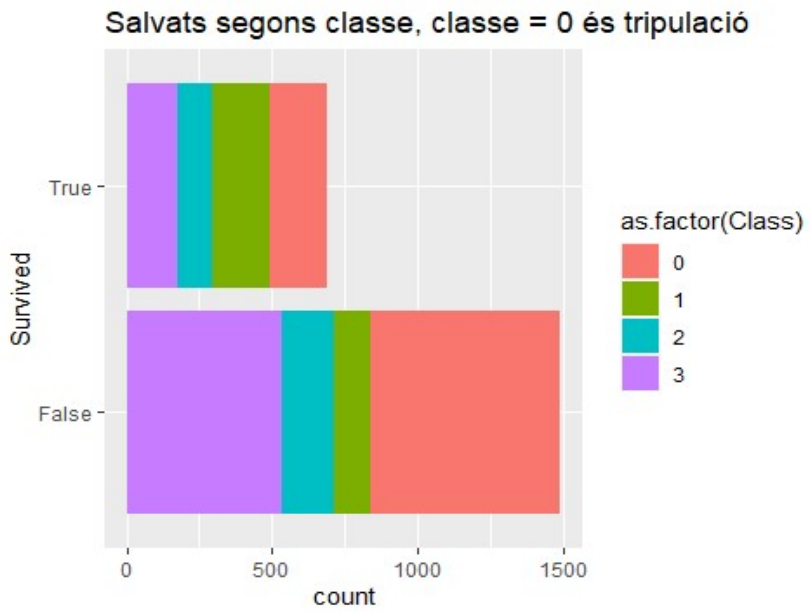
4- Salvats segons els grups



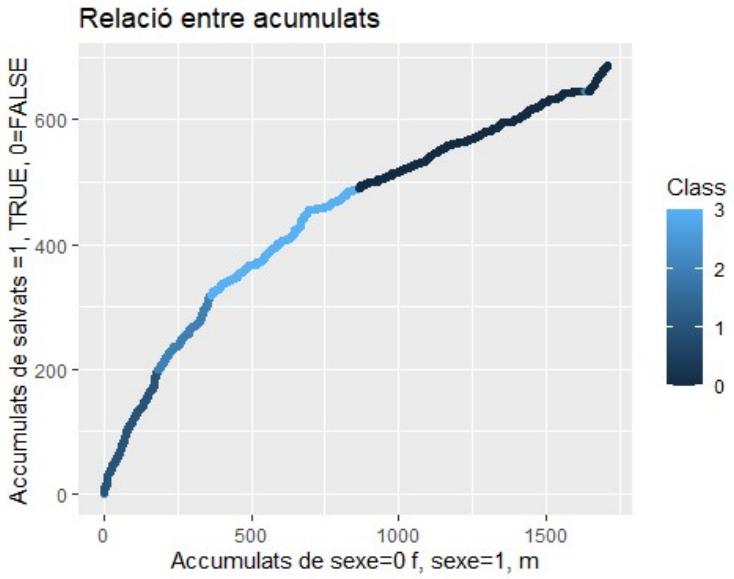
5- Nombre de salvats per cada sexe



6- Nens salvats segons la classe



7- Salvats de cada classe i de la tripulació



8- Com canvia la relació de salvats segons el sexe d'acord amb les classes

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

5.2 Taules

Algunes taules són esteses i queden retallades, però es poden comprovar en R-Studio amb el RMD.

- Taula de contingència de % del total de supervivents

```
##       False     True
##   f 28.35821 71.64179
##   m 79.43761 20.56239
```

- Salvats, en %, segons els grups:

```
##                    False      True
##   1c0ser         42.400000  57.600000
##   1c1serv        21.428571  78.571429
##   1c2serv        33.333333  66.666667
##   1classServ     29.268293  70.731707
##   2class         57.553957  42.446043
##   3class         75.564972  24.435028
##   Deck           30.508475  69.491525
##   Engineering    78.683386  21.316614
##   garantee      100.000000   0.000000
##   officer        50.000000  50.000000
##   Orchestra     100.000000   0.000000
##   Postman       100.000000   0.000000
##   restaurant     95.588235   4.411765
##   Victualling    79.177378  20.822622
```

1c0ser, passatger de primera classe sense servent
1c1ser, passatger de primera classe amb 1 servent també en 1a
1c2ser, passatger de primera classe amb 2 servents també en 1a
1classServ, servent en 1a classe (secretary, valet, nurse,..)
2class, passatger de segona classe
3class, passatger de tercera classe
resta: tripulació (deck, marineria, engineering, tripulació a càrrec de les màquines,...

- Nens menors de 15 anys salvats i ofegats, ordenats per grup

```
##   Class Survived n_nens
## 1     1 False       1
## 2     1 True        6
## 3     2 False       1
## 4     2 True       26
## 5     3 False      63
## 6     3 True       26
```

- Salvats: Ocupació dels bots:

```
##    Lifeboat Survived n_by_boat
##  1 ""       False        1489
##  2 "1"      True           18
##  3 "10"     True           31
##  4 "11"     True           46
##  5 "12"     True           20
##  6 "13"     True           65
##  7 "14"     True           44
##  8 "15"     True           59
##  9 "16"     True           35
## 10 "2"      True           20
```

Els False són ofegats.

- Com de plens, en %, anaven el bots ?

```
##    Lifeboat Survived n_by_boat capacity percentFull
##  1 ""       False        1489       NA          NA
##  2 "1"      True           18       65        27.7
##  3 "10"     True           31       65        47.7
##  4 "11"     True           46       65        70.8
##  5 "12"     True           20       65        30.8
##  6 "13"     True           65       65        100
##  7 "14"     True           44       65        67.7
##  8 "15"     True           59       40        148.
##  9 "16"     True           35       40        87.5
## 10 "2"      True           20       65        30.8
## # … with 11 more rows
```

- El cas del 100 % percentatge de salvats dels que abordaren el Titanic en Cherbourg

```
             False      True
1c0ser       38.00000  62.00000
1c1serv       0.00000 100.00000
1c2serv      33.33333  66.66667
1classServ   15.00000  85.00000
2class       40.90909  59.09091
3class       67.67677  32.32323
```

- Llistat del grup de 1a classe, amb un servent també en 1a classe, que se salvaren tots

X	Name	Age
<int>	<fctr>	<dbl>
293	Aubart, Mrs. Léontine Pauline	24
294	Bucknell, Mrs. Emma Eliza (née Ward)	59
295	Cardeza, Mrs. Charlotte Wardle (née Drake)	58
296	Cardeza, Mr. Thomas Drake Martinez	36
299	Douglas, Mrs. Mahala (née Dutton)	48
300	Duff Gordon, Lucy Christiana, Lady (née Sutherland)	48
302	Harper, Mr. Henry Sleeper	48
307	Peñasco y Castellana, Mrs. Maria Josefa (née Perez de Soto y Vallejo)	22
309	Ryerson, Mrs. Emily Maria (née Borie)	48
310	Ryerson, Master John Borie "Jack"	13
311	Spedden, Mrs. Margaretta Corning (née Stone)	39
312	Spedden, Master Robert Douglas	6
313	Spencer, Mrs. Marie Eugénie (née Demougeot)	45
317	Thayer, Mrs. Marian Longsteth (née Morris)	39

- Llistat del grup dels servents de 1a classe del grup anterior

X	Name	Age	Boarded	Position	Lifeboat	Bo
<int>	<fctr>	<dbl>	<fctr>	<fctr>	<fctr>	<fc
4	Miss Rosalie Bidois	46	C	maid	4	
5	Miss Caroline Louise Endres	39	C	nurse	4	
6	Miss Emma Säugesser	25	C	maid	9	
7	Miss Albina Bazzani	36	C	maid	8	
8	Miss Annie Moore Ward	38	C	maid	3	
9	Mr. Gustave J. Lesueur	35	C	valet	3	
14	Miss Berthe Leroy	27	C	maid	3	
15	Miss Laura Mabel Francatelli	31	C	secretary	1	
17	Mr. Victor Giglio	24	C	valet		

Se salven 8 de 9.

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

6.1 Dades dels salvats

Ja havíem dit que 'sabíem que no tots els grups es van salvar per igual'. La idea és analitzar en les dades aquesta qüestió, i si és possible, el perquè.

Reportem els 5 primers grups de Group, per percentatge de salvats:

- Salvats, en %, segons els grups:

```
##                 False      True
##   1c1serv      21.428571 78.571429
##   1classServ   29.268293 70.731707
##   Deck         30.508475  69.491525

##   1c2serv      33.333333 66.666667

##   1c0ser       42.400000  57.600000
```

1c0ser, passatger de primera classe sense servent
1c1ser, passatger de primera classe amb 1 servent també en 1a
1c2ser, passatger de primera classe amb 2 servents també en 1a
1classServ, servent en 1a classe (secretary, valet, nurse,..)
deck, marineria

Els passatgers de 1a classe amb un servent són el grup que més se salva (78.57%), seguit pels mateixos servents acompanyants de 1a classe (70.73%).

El següent és la tripulació de 'deck', la marineria, (69,49%). Després la 1a classe amb 2 servents acompanyants (66,66%).

Respecte als menors de quinze anys, hi ha més ofegats en 3a classe (63) que tots els que es van salvar:

- Nens menors de 15 anys salvats i ofegats, ordenats per grup

```
##   Class Survived n_nens
## 1     1 False       1
## 2     1 True        6
## 3     2 False       1
## 4     2 True       26
## 5     3 False      63
## 6     3 True       26
```

Tres bots anaven al 30%, o menys, de capacitat. La resta no anaven plens. (Bots 1 a 14 caben 65, bots 15 i 16 caben 40, els bots plegables A, B, C i D, 47). False són ofegats. Quin percentatge de ple anaven els bots? La resposta és molt variable: l'1 anava al 27,7% mentre que el 15 sobrecarregat al 147,5%

- Com de plens anaven el bots ?

```
##    Lifeboat Survived n_by_boat capacity percentFull
##  1 ""       False        1489       NA          NA
##  2 "1"      True           18       65        27.7
##  3 "10"     True           31       65        47.7
##  4 "11"     True           46       65        70.8
##  5 "12"     True           20       65        30.8
##  6 "13"     True           65       65        100
##  7 "14"     True           44       65        67.7
##  8 "15"     True           59       40        148.
##  9 "16"     True           35       40        87.5
## 10 "2"      True           20       65        30.8
## # … with 11 more rows
```

Hi ha molts grups ofegats el 100%. Hi ha algun grup salvat al 100%? Sí.

- Percentatge de salvats dels que abordaren el Titanic en Cherbourg

```
             False      True
1c0ser       38.00000  62.00000
1c1serv       0.00000 100.00000
1c2serv      33.33333  66.66667
1classServ   15.00000  85.00000
2class       40.90909  59.09091
3class       67.67677  32.32323
```

Per tant, és el group dels 14 passatgers de 1a classe, amb un servent també en 1a classe, que abordaren al principi del viatge inaugural del Titanic. Els seus servents també en 1a classe se salvaren 8 de 9, el 88,9%:

6.2 Predicció

En la versió reduïda:

Amb un model d'arbre de decisió, quines són les regles principals que porten a sobreviure o no? L'arbre de decisió no classifica bé a soles un 21,8%. D'acord amb el model,

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

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

Potser per haver-hi pocs casos, que no hi ha cap regla respecte a si els nens se salvin més que els adults. Hem fet un test d'hipòtesis, de dues mostres independents, sobre la mitjana d'edat dels que es van salvar i els que no. Ens dóna un p-value de 0.7078 >0.05 amb el resultat 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?

Hem creat el següent model de regressió lineal:

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

amb un coeficient de determinació molt satisfactori de 0.996, i podem predir el nombre de supervivents amb la resta de variables acumulades.

7-Conclusions

S'ha aconseguit crear un fitxer únic amb 2178 dades de la Wikipedia en 11 columnes, net i preparat per l'anàlisi de dades.

S'ha aconseguit l'objectiu de l'anàlisi de trobar els 'Titanic facts', el primer i més important és el percentatge de salvats per grups. Hi havia dades molt generals d'aquest percentatge que podem confirmar, així com molts nous que hem aportat a la llum de les dades.

Un fet és l'extrema desigualtat dels grups amb relació a aquest percentatge de salvats. Els bots no s'aprofitaren bé sinó que alguns partiren aviat i no anaven plens, com demostren les dades.

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, com garantee, orchestra i postman.

La terrible quantitat de nens ofegats en la 3a classe.

Totes aquestes dades no les hem trobat en la xarxa, sinó que són resultats de l'anàlisi.

Hem resolt les preguntes amb anàlisi de les dades, com les reportades en l'apartat 7, amb tests variats en les dades, amb models de regressió, arbres de decisió, taules de contingència, comparació de mitjanes amb diversos tests paramètrics i no paramètrics, predicció segons dos models de l'apartat 4, i la neteja de dades i verificació de normalitat i homoscedasticitat de les variables adients, en els apartats 3 i 4.

Hem reportat gràfiques diverses que fan visibles els resultats més destacats. Seria massa llarg reflectir ací tots els resultats, que estan detallats en els rmd i html que s'adjunten.

Contribucions	*Noms*
Investigació prèvia	Antonio i Alexandre
Redacció de les respostes	Antonio i Alexandre
Desenvolupament del codi	Antonio i Alexandre