Universitat Oberta de Catalunya

Estudis d'Informàtica, Multimedia i Telecomunicacions

Tipologia i cicle de vida de les dades: Pràctica 2 - Neteja i anàlisi de les dades

Autor: Albert Eduard Merino Pulido

Gener 2020

1 Titanic Data Analysis

1.1 Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes.

1.2 Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat 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.

1.3 Objectius

Els objectius concrets d'aquesta pràctica són:

- 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 (integració, neteja i validació) 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 les habilitats d'aprenentatge que els permetin continuar estudiant d'unam anera que haurà de ser en gran manera autodirigida o autònoma.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

1.4 Descripció de la PAC a realitzar

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i justificar) són les següents:

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?
2. Integració i selecció de les dades d'interès a analitzar.
3. Neteja de les dades.
    1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?
    2. Identificació i tractament de valors extrems.
4. Anàlisi de les dades.
    1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).
    2. Comprovació de la normalitat i homogeneïtat de la variància.
    3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.
5. Representació dels resultats a partir de taules i gràfiques.
6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

---

2 Resolució

Important: La resolució de la pràctica es pot veure en el següent enllaç https://data-science-env.github.io/Titanic-Data-Analysis/

2.1 Descripció del dataset

En aquesta pràctica es treballarà amb el ben conegut dataset Titanic. Aquest dataset conté dades de l'històric enfonsament del vaixell Titanic. És un dataset molt utilitzat i treballat per molta gent de tot el món per la seva estructura simple per a iniciar-se al món del machine learning. A més cal destacar que les dades que tracta poden entendres facilment. El dataset proporciona informació sobre el destí dels passatgers al Titànic, resumits segons l'estat econòmic (classe), el sexe, l'edat i la supervivència. L'objectiu típic és desenvolupar models capaços de fer prediccions sobre si un passatge va sobreviure o no a l'accident.

Els fitxers necessaris han estat descarregats des de la pàgina web Kaggle mitjançant el següent enllaç. Com podem veure Kaggle ens proporciona dos fitxers anomenats train i test. En el primer es troben les dades d'entrenament, així doncs inclou una columna indicant si el passatger sobreviu o no. En canvi, en el fitxer de test no existeix cap columna que indiqui si el passatger va sobreviure. Això ve donat així, ja que Kaggle fa una competició amb els models generats pels diferents participants. A partir del model generat amb el dataset d'entrenament, utilitzen el dataset de test per avaluar el model i així generar una puntuació indicant com de bé és el model.

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

Inicialment caldrà que carreguem les dades d'entrenament que utilitzarem per fer l'anàlisi. El fitxer proporcionat és un fitxer en format CSV (Comma Separated Values). A la càrrega de dades indicarem que no volem que es transformin automàticament totes les variables de tipus caràcter a factors, això serà una decisió que prendrem a través de l'anàlisi.

```
# Carreguem els paquets R que utilitzarem
library(dplyr)
library(ggplot2)

# Guardem el joc de dades train en una variable
data <- read.csv('../data/titanic-train.csv', stringsAsFactors = FALSE)

# Verifiquem l'estructura del joc de dades
str(data)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)
" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Com podem veure el dataset Titanic té un total de 891 observacions en el fitxer d'entrenament. A més podem veure que compta amb 12 variables de les quals hi ha 7 de contínues i 5 de categòriques, les quals descrivim a continuació:

- `PassengerId` : Identificador únic per a cada passatger
- `Survived` : Indica si el passatger sobreviu al naufragi (0 = No , 1 = Si)
- `Pclass` : Classe del passatger (1 = primera, 2 = segona, 3 = tercera)
- `Name` : Nom del passatger
- `Sex` : Sexe del passatger
- `Age` : Edat en anys del passatger
- `SibSp` : Nombre de germans/germanes/germanastres/germanastres a bord
- `Parch` : Nombre de pares/fills a bord
- `Ticket` : Número de bitllet
- `Fare` : Preu pagat pel bitllet
- `Cabin` : Identificador de la cabina assignada
- `Embarked` : Port on va embarcar el passatger

2.3 Neteja de les dades

Treballem els atributs amb valors buits.

```
# Estadístiques de valors buits
colSums(is.na(data))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

```
colSums(data=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

```
# Prenem valor "C" per als valors buits de la variable "Embarked"
data$Embarked[data$Embarked==""]="C"

# Prenem la mitjana per a valors buits de la variable "Age"
data$Age[is.na(data$Age)] <- mean(data$Age,na.rm=T)
```

Discretitzem quan té sentit i en funció de cada variable. Per decidir quines variables discretitzar primerament veurem quants registres diferents existeixen en cada variable. Un cop generada aquestà taula elegirem aquelles variables que continguin pocs registres diferents.

```
# Per a quines variables tindria sentit un procés de discretització?
apply(data,2, function(x) length(unique(x)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##         891           2           3         891           2          89
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           7           7         681         248         148           3
```

```
# Discretitzem les variables amb poques classes
cols<-c("Survived","Pclass","Sex","Embarked")
for (i in cols){
  data[,i] <- as.factor(data[,i])
}

# Després dels canvis, analitzem la nova estructura del joc de dades
str(data)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)
" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
```

Les variables discretitzades han estat `Survived` amb 2 classes, `Pclass` amb 3 classes, `Sex` amb 2 classes i `Embarked` amb 3 classes. També podem veure com la variable `PassangerId` efectivament conté un valor diferent en cada observació com era d'esperar.

Per últim podem treballar els valors 0.

```
# Estadístiques de valors 0
colSums(data==0)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         549           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##         608         678           0          15           0           0
```

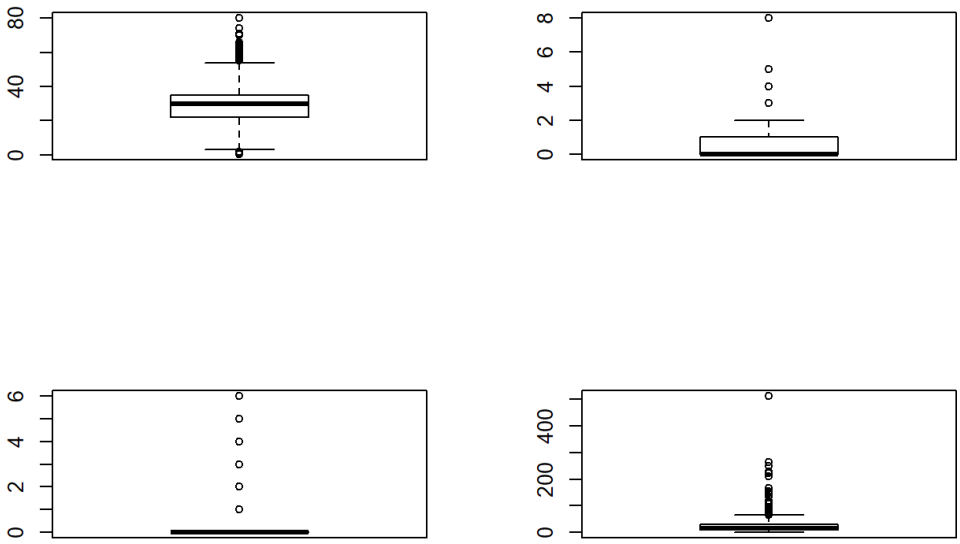
Encara que hi ha 4 variables amb observacions amb valor 0, no considerarem aquest valor com a dolent a causa del tipus de variables que estem tractant. No obstant podem observar amb més detall les observacions on la variable `Fare` conté el valor 0 ja que hi ha poques observacions. D'aquestes dades les conclusions que podem treure és que els passatgers que aparentment van viatjar de franc eren tot homes, tots van embarcar al mateix port i no anaven acompanyats de ningun familiar.

```
# Estadístiques de valors 0
data[data$Fare==0,]
```

```
##     PassengerId Survived Pclass                            Name  Sex      Age
## 180         180        0      3              Leonard, Mr. Lionel male 36.00000
## 264         264        0      1            Harrison, Mr. William male 40.00000
## 272         272        1      3     Tornquist, Mr. William Henry male 25.00000
## 278         278        0      2      Parkes, Mr. Francis "Frank" male 29.69912
## 303         303        0      3  Johnson, Mr. William Cahoone Jr male 19.00000
## 414         414        0      2   Cunningham, Mr. Alfred Fleming male 29.69912
## 467         467        0      2            Campbell, Mr. William male 29.69912
## 482         482        0      2 Frost, Mr. Anthony Wood "Archie" male 29.69912
## 598         598        0      3              Johnson, Mr. Alfred male 49.00000
## 634         634        0      1    Parr, Mr. William Henry Marsh male 29.69912
## 675         675        0      2       Watson, Mr. Ennis Hastings male 29.69912
## 733         733        0      2             Knight, Mr. Robert J male 29.69912
## 807         807        0      1           Andrews, Mr. Thomas Jr male 39.00000
## 816         816        0      1                Fry, Mr. Richard male 29.69912
## 823         823        0      1  Reuchlin, Jonkheer. John George male 38.00000
##     SibSp Parch Ticket Fare Cabin Embarked
## 180     0     0   LINE    0              S
## 264     0     0 112059    0   B94        S
## 272     0     0   LINE    0              S
## 278     0     0 239853    0              S
## 303     0     0   LINE    0              S
## 414     0     0 239853    0              S
## 467     0     0 239853    0              S
## 482     0     0 239854    0              S
## 598     0     0   LINE    0              S
## 634     0     0 112052    0              S
## 675     0     0 239856    0              S
## 733     0     0 239855    0              S
## 807     0     0 112050    0   A36        S
## 816     0     0 112058    0  B102        S
## 823     0     0  19972    0              S
```

A continuació donem una ullada als outliers o valors atípics. Aquests valors solen representar valors extrems dins d'un conjunt de dades. Normalment se sol considerar aquests valors com els valors que marxen com a mínim dues desviacions estàndard dins del conjunt de dades. Per tal d'identificar-los farem ús del mètode boxplot() proporcionat per R el qual ens mostrarà gràficament els outiliers en cas que existeixin.

```
par(mfrow=c(2,2))
boxplot(data$Age)
boxplot(data$SibSp)
boxplot(data$Parch)
boxplot(data$Fare)
```



En aquest cas s'han escollit únicament aquelles variables numèriques i no factoritzades. En el plot de dalt a l'esquerra podem veure com l'edat dels passatges ronda els trenta anys. Els valors extrems que ens apareixen són nadons i padrins que no superen edats fora del normal, així doncs s'han de considerar tots els casos. El plot de dalt a la dreta i baix a l'esquerra no es qüestionaran, ja que és possible que surtin els resultats donats, ja que poden haver-hi famílies a bord. Per últim, en plot de baix a l'esquerra sí que veiem que hi ha un valor que marxa massa de la resta, per tant aquest sí que l'estudiarem més endavant per veure les característiques associades a la seva observació.

2.4 Anàlisi de les dades

Abans de començar a analitzar les dades crearem una nova variable que anomenarem `NumFamiliars` on tindrem el nombre de familiars a bord de cada passatger i per tant poder fer analisis per nombre de familiars.

```
# Construïm un atribut nou: NumFamiliars
data$NumFamiliars <- data$SibSp + data$Parch + 1;

# Després dels canvis, analitzem la nova estructura del joc de dades
str(data)
```

```
## 'data.frame':    891 obs. of  13 variables:
##  $ PassengerId : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived    : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass      : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name        : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer
)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex         : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age         : num  22 38 26 35 35 ...
##  $ SibSp       : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch       : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket      : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare        : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin       : chr  "" "C85" "" "C123" ...
##  $ Embarked    : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ NumFamiliars: num  2 2 1 2 1 1 1 5 3 2 ...
```

Com haviem dit durant la neteja de dades estudariem el cas atipic on s'ha pagat una gran quantitat pel bitllet. La quantitat pagada ha estat 512.3292 lliures, i aquesta l'han pagada tres passatgers. Per tant comprovem si aquests van pujar junts i per tant poden ser familia.

```
fare_outliers <- boxplot.stats(data$Fare)$out
fare_outliers[fare_outliers>300]
```

```
## [1] 512.3292 512.3292 512.3292
```

```
data[data$Fare>500,]
```

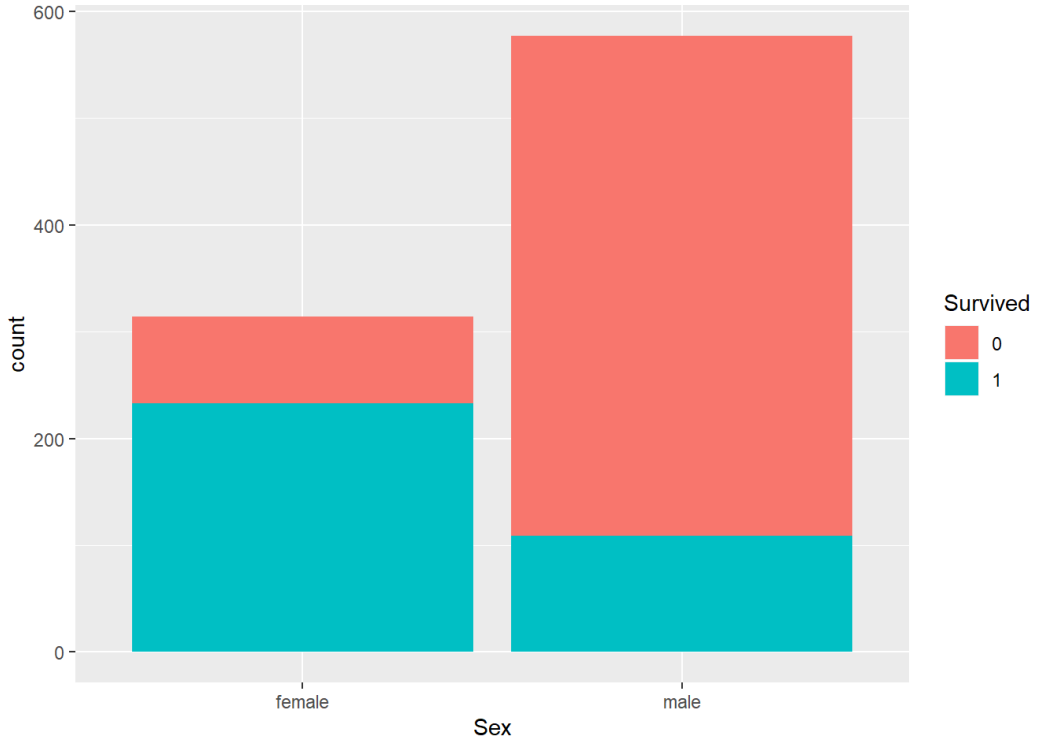
```
##     PassengerId Survived Pclass                                Name    Sex Age
## 259         259        1      1                    Ward, Miss. Anna female  35
## 680         680        1      1 Cardeza, Mr. Thomas Drake Martinez   male  36
## 738         738        1      1              Lesurer, Mr. Gustave J   male  35
##     SibSp Parch    Ticket     Fare       Cabin Embarked NumFamiliars
## 259     0     0 PC 17755 512.3292                   C            1
## 680     0     1 PC 17755 512.3292 B51 B53 B55       C            2
## 738     0     0 PC 17755 512.3292        B101       C            1
```

Podem veure que els tres passatgers sembla que no tenen cap relació familiar entre ells. Tots tres ronden la mateixa edat, van embarcar al mateix port i viatgen en primera classe. El primer i el tercer passatger no tenen ni germans ni pares/fills així doncs podem concloure que viatgen sols. En canvi el segon passatger podem veure com sí que té un familiar a bord. Així doncs podem concloure que els tres individus, com a mínim, no tenen cap relació familiar entre ells i per tant el preu pagat pel bitllet podria ser simplement per un possible estatus social alt.

A continuació, ens proposem analitzar les relacions entre les diferents variables del joc de dades i la variable objectiu `Survived`.

En aquest primer plot podem veure la relació entre la variable `Sex` i `Survived`. Podem veure com en tant per cent les dones van sobreviure més que els homes, això ens pot fer entendre que es va prioritzar el salvament de les dones. Dincs del grup de les dones podem veure com un 74.2% d'elles van sobreviure. Si en canvi fem la mateixa visualització per als homes, podem veure com solament el 18.9% dels homes va sobreviure.

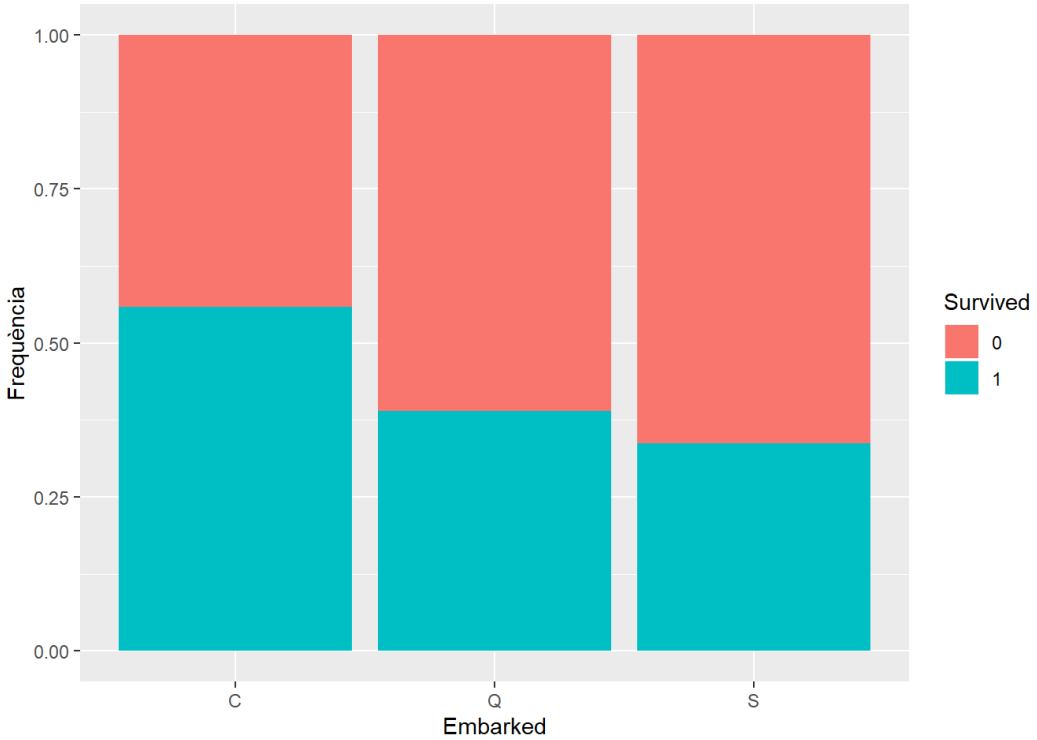
```
# Visualitzem la relació entre les variables "sex" i "survival":
ggplot(data=data,aes(x=Sex,fill=Survived))+geom_bar()
```




```
t<-table(data$Sex,data$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##                 0        1
##   female 25.79618 74.20382
##   male   81.10919 18.89081
```

```
# Un altre punt de vista. Survival com a funció de Embarked:
ggplot(data=data, aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+ylab("Frequència")
```



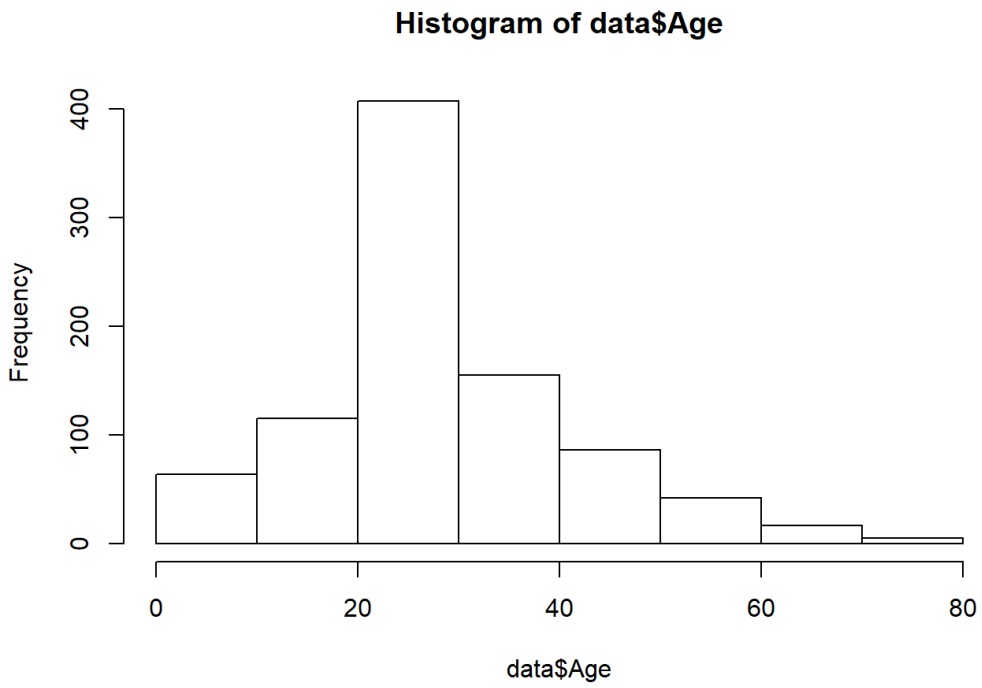
Obtenim una matriu de percentatges de freqüència. Veiem, per exemple que la probabilitat de sobreviure si es va embarcar en “C” és d’un 55,88%

```
t<-table(data$Embarked,data$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##            0        1
##   C 44.11765 55.88235
##   Q 61.03896 38.96104
##   S 66.30435 33.69565
```

A continuació anem a veure en el següent histograma com està distribuïda la població en funció de la seva edat. Amb això es vol comprovar si es va tenir preferència per salvar als nens, considerats fins als catorze anys, abans que els adults, considerats de 15 fina el valor màxim. Podem veure com l’edat de la majoria dels passatgers oscil·la entre els 20 i els 30 anys.

```
hist(data$Age)
```



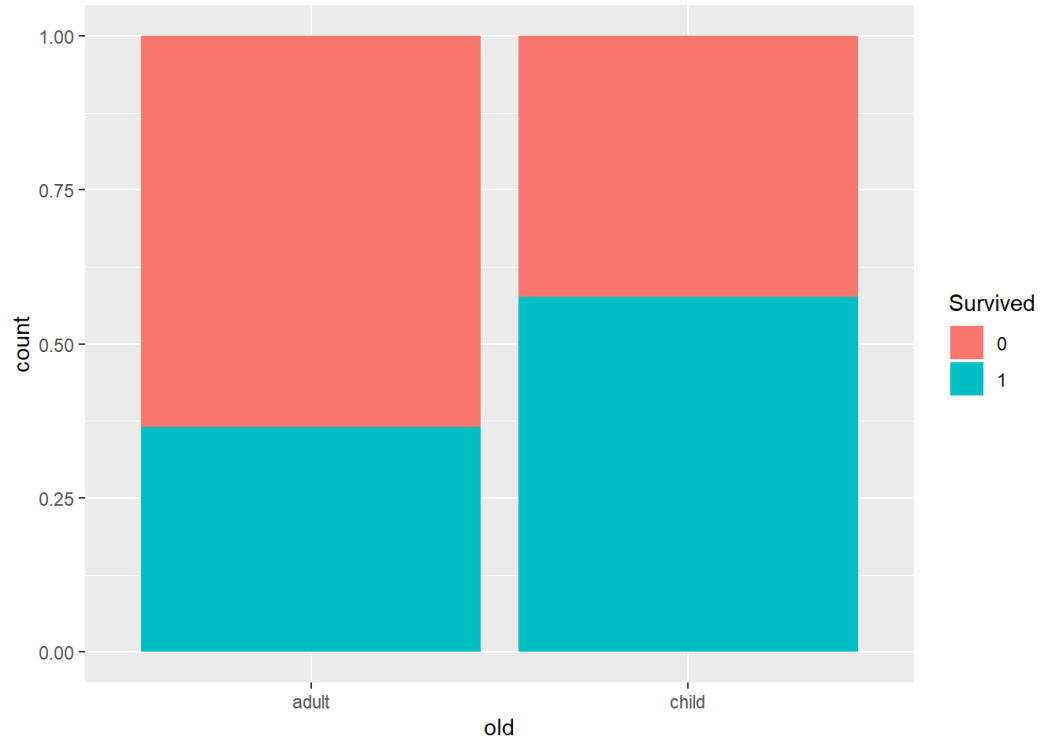
A continuació anem a generar un variable més per al dataset indicant si el passatger és un nen o un adult, d'aquesta manera podrem generar grafics que ens mostrin la hipotesis que estudiem.

```
data$old <- 'adult'
data$old[data$Age < 15]<-'child'

# Comprovem quants nens i adults tenim en la població
table(data$old)
```

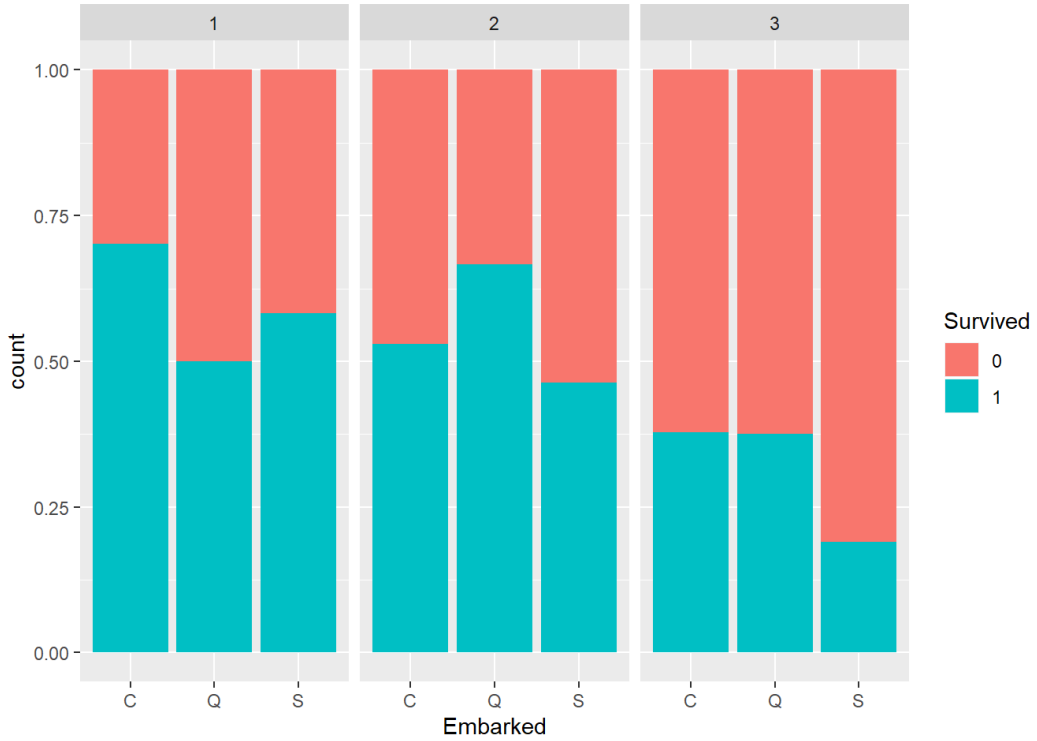
```
##
## adult child
##   813    78
```

```
# Visualitzem la relació entre les variables "old" i "survival":
ggplot(data = data,aes(x=old,fill=Survived))+geom_bar(position="fill")
```



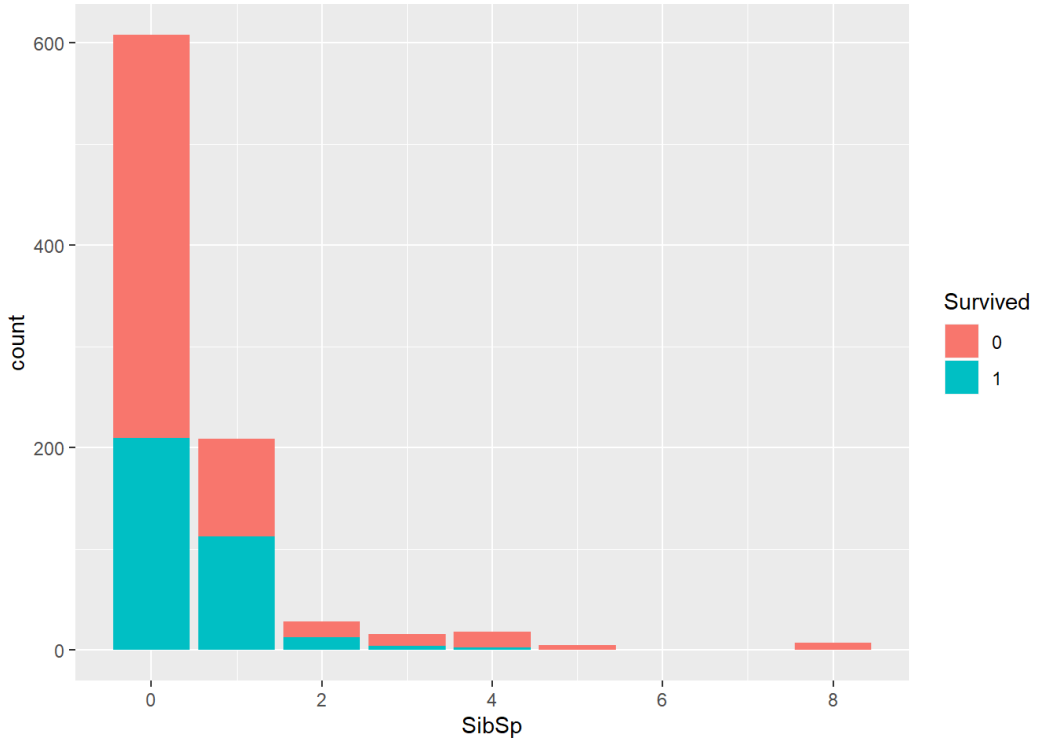
A continuació generarem un gràfic de freqüències on representarem 3 variables: `Embarked`, `Survived` i `Pclass`. Amb aquest plot podem veure com els passatgers que anaven en primera o segona van sobreviure més que els que els passatgers que anaven en tercera classe.

```
# Mostrem el gràfic d'embarcats per Pclass
ggplot(data = data,aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+facet_wrap(~Pclass)
```

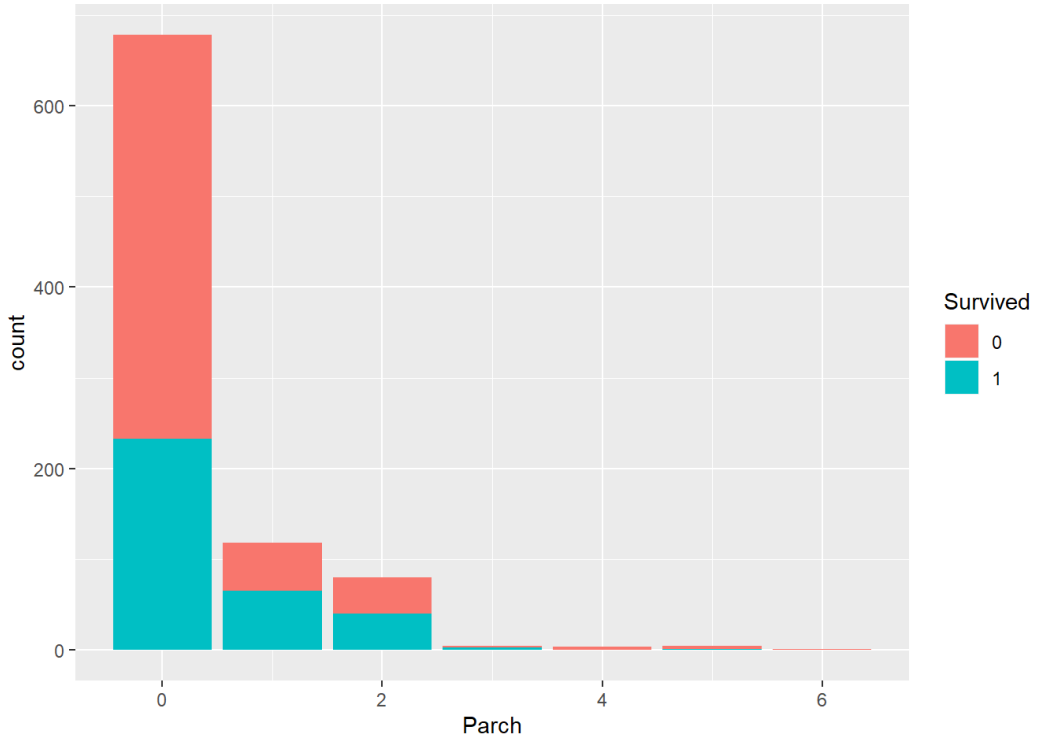


A continuació comparem amb dos gràfics de freqüències les parelles de variables Survived-SibSp i Survived-Parch. Veiem com la forma d'aquests dos gràfics és similar. Aquest fet ens pot indicar presència de correlacions altes.

```
ggplot(data = data,aes(x=SibSp,fill=Survived))+geom_bar()
```

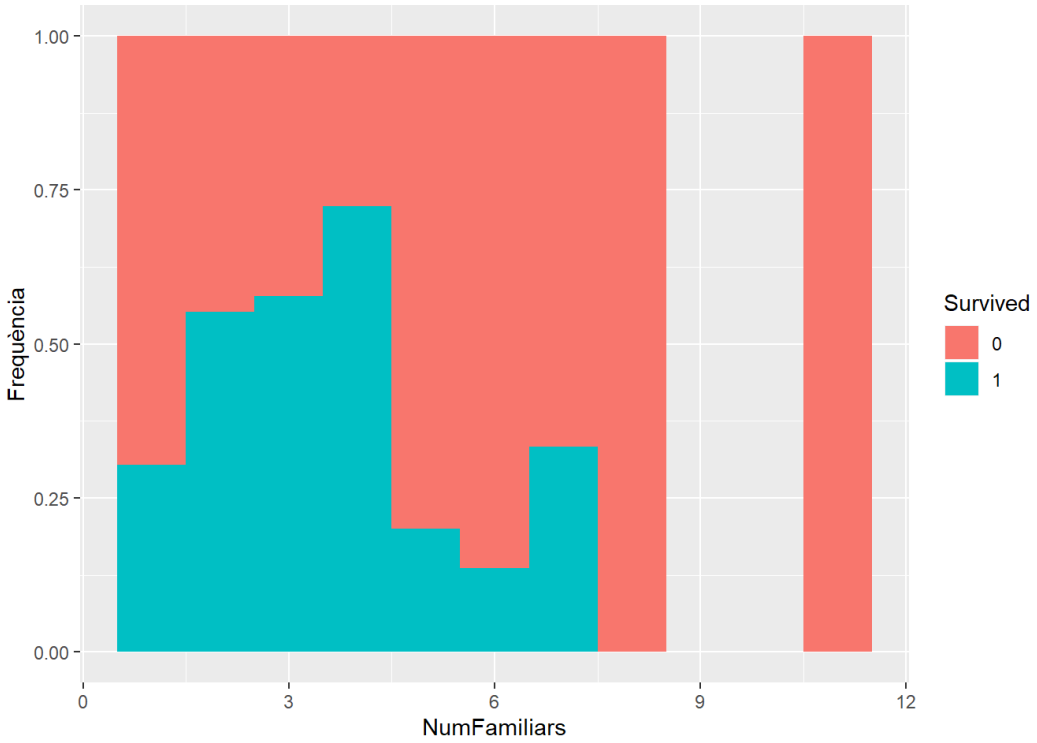



```
ggplot(data = data,aes(x=Parch,fill=Survived))+geom_bar()
```



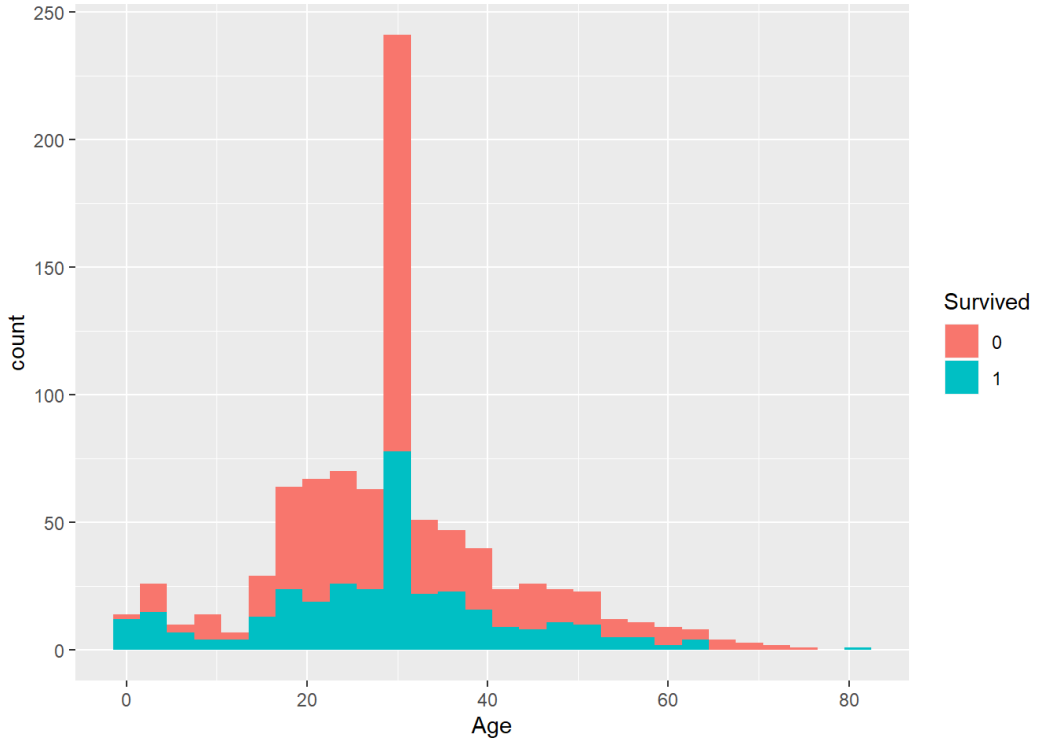
A partir de la variable `NumFamiliars` comprovem quina ha estat la supervivència en funció del nombre de familiars a bord. Podem observar com les famílies de 2 a 4 membres tenen més del 50% de possibilitats de supervivència. Podem veure que a partir de 5 membres la supervivència és molt escassa fins al punt de trobar que en famílies de 8 o 11 membres la supervivència és nul·la.

```
ggplot(data =
data[!is.na(data$NumFamiliars),],aes(x=NumFamiliars,fill=Survived))+geom_histogram(binwidth =1,positio
n="fill")+ylab("Frequència")
```

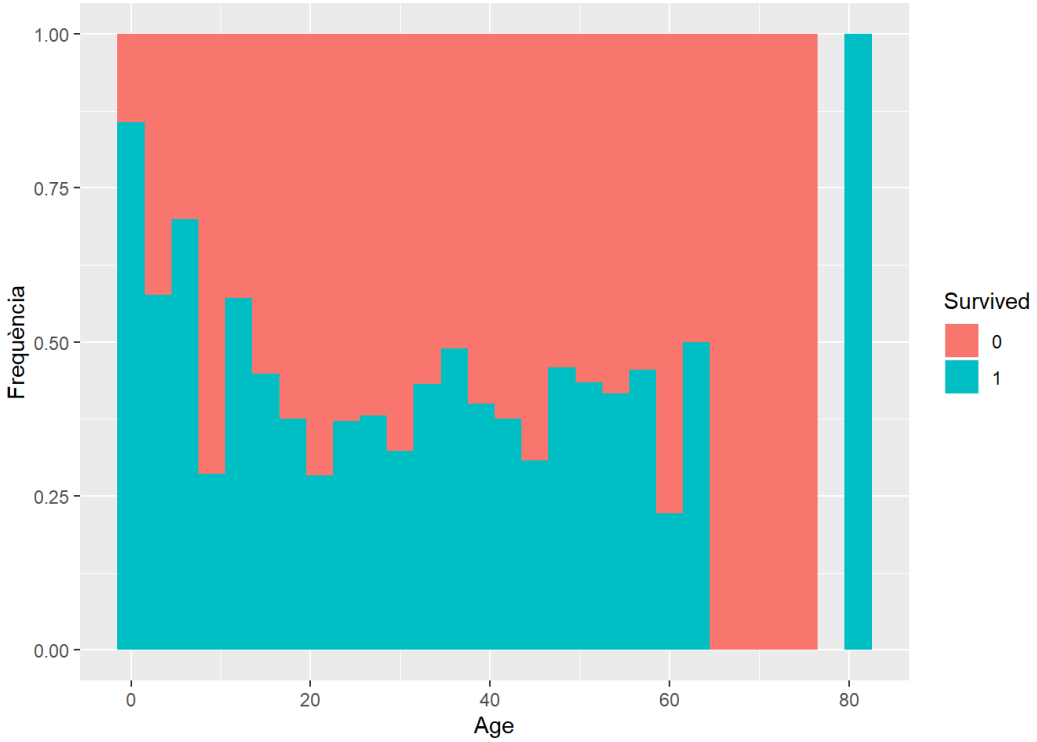


Vegem ara dos gràfics que ens compara els atributs Age i Survived. Observem com el paràmetre position="fill" ens dóna la proporció acumulada d'un atribut dins d'un altre

```
# Survival com a funció de age:
ggplot(data = data[!(is.na(data$Age)),],aes(x=Age,fill=Survived))+geom_histogram(binwidth =3)
```




```
ggplot(data = data[!is.na(data$Age),],aes(x=Age,fill=Survived))+geom_histogram(binwidth = 3,position="
fill")+ylab("Freqüència")
```



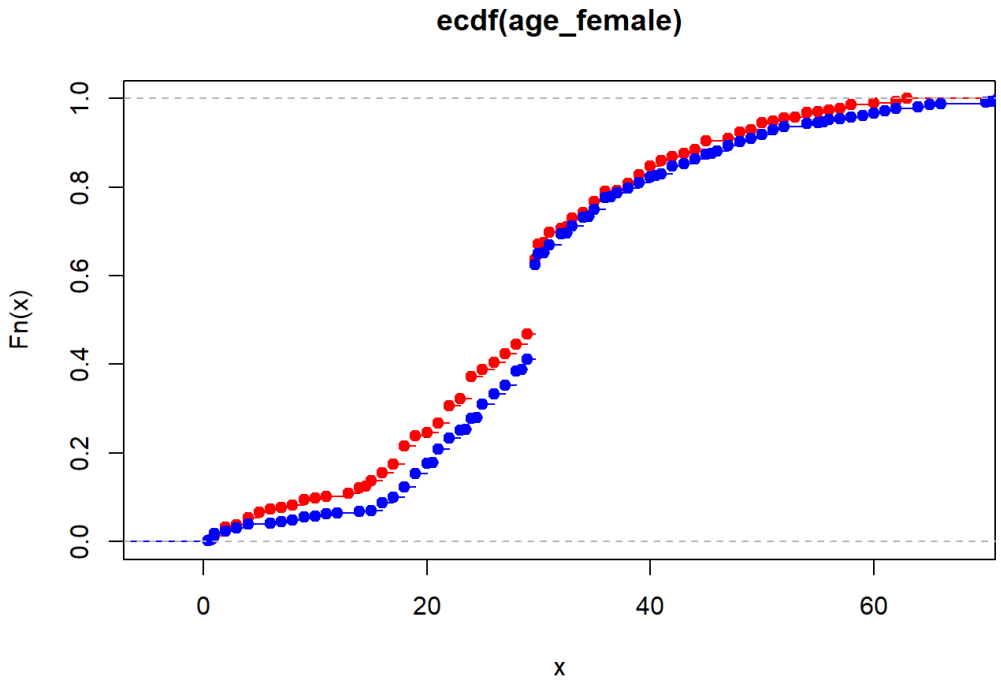
A continuació és farà un estudi utilitzant anàlisis estadística. Aquest estudi es farà sobre la variable `Age` comparant les edats dels homes i de les dones a traves de la variable `Sex`. Em pogut veure en aquest ultim plot com la freqüencia de supervivencia està aprop del 50%. Així doncs amb aquest analisi volem veure si es segueix mantenint aquest percentatge quan separem homes i dones en dos grups.

Amb l'objectiu de verificar la suposició de la normalitat, algunes de les proves més habituals són els tests de Kolmogorov-Smirnov i de Shapiro-Wilk. Tot i que tots dos comparen la distribució de les dades amb una distribució normal, el test de Shapiro-Wilk es considera un dels mètodes més potents per contrastar la normalitat. Assumint com a hipòtesi nul·la que la població està distribuïda normalment, si el p-valor és més petit que el nivell de significació, generalment α=0,05, llavors la hipòtesi nul·la és rebutjada i es conclou que les dades no compten amb una distribució normal. Si, per contra, el p-valor és major a α, es conclou que no es pot rebutjar aquesta hipòtesi i s'assumeix que les dades segueixen una distribució normal.

El següent fragment de codi a R mostra com es poden aplicar aquestes proves, mitjançant les funcions ks.test() i shapiro.test(), respectivament. A més és fara ús del mètode ecdf() per mostrar gràficament la Empirical Cumulative Distribution Function, veient d'aquesta manera una comparativa entre les edats dels homes i les dones al llarg del dataset.

Podem veure com ambudes linies tenen la mateixa forma, així doncs podem dir que tenim un dataset compensat en la variable `Age`. Si ens fixem en els resultat de ks i shapiro veiem com el p-value està per sota del llindar α, així doncs sembla que no segueix una distribució normal.

```
age_female <- data$Age[data$Sex=="female"]
age_male <- data$Age[data$Sex=="male"]
plot(ecdf(age_female),col="red")
lines(ecdf(age_male),col="blue")
```




```
ks.test(age_female,age_male)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  age_female and age_male
## D = 0.095315, p-value = 0.04971
## alternative hypothesis: two-sided
```

```
shapiro.test(data$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Age
## W = 0.95882, p-value = 3.969e-15
```

```
var.test(age_female,age_male)
```

```
##
##  F test to compare two variances
##
## data:  age_female and age_male
## F = 0.97982, num df = 313, denom df = 576, p-value = 0.8453
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8085848 1.1940983
## sample estimates:
## ratio of variances
##          0.9798199
```

En la prova t de Student, la hipòtesi nul·la assumeix que les mitjanes dels grups de dades són les mateixes, mentre que en les proves no paramètriques s'assumeix que les distribucions dels grups de dades són les mateixes. Per tant, només si el p-valor resultant de la prova és menor al nivell de significació es rebutjarà la hipòtesi nul·la i es conclourà que hi ha diferències estadísticament significatives entre els grups de dades analitzades.

```
t.test(age_female,age_male)
```

```
##
##  Welch Two Sample t-test
##
## data:  age_female and age_male
## t = -2.5257, df = 648.52, p-value = 0.01179
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.0688028 -0.5093856
## sample estimates:
## mean of x mean of y
##  28.21673  30.50582
```

```
t.test(age_female,age_male,alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  age_female and age_male
## t = -2.5257, df = 648.52, p-value = 0.005893
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.7961705
## sample estimates:
## mean of x mean of y
##  28.21673  30.50582
```

```
t.test(age_female,age_male,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  age_female and age_male
## t = -2.5257, df = 648.52, p-value = 0.9941
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.782018       Inf
## sample estimates:
## mean of x mean of y
##  28.21673  30.50582
```

Finalment, seguint els criteris de la pràctica es generarà de nou un CSV amb el dataset al final de l'anàlisi.

```
write.csv(data,"../data/titanic-end.csv", row.names = FALSE)
```

2.5 Conclusions

Una vegada conclos l'estudi presentem les següents conclusions sobre el dataset de titanic. L'estudi s'ha centrat en veure la relació que hi ha hagut entre els passatgers del vaixell i si van sobreviure a l'accident.

S'ha pogut observar com les dades són molt netes, això és degut a que aquest dataset esta pensat per ser estudiat per gent principiant al món del Data Science. S'ha pogut comprovar com solament havien valors buits per a la variable `Age` els quals han estat emplenats amb el valor mitja de la població. També s'ha comprovat mitjançant diagrames de caixa l'existencia de valors extrems en les variables `Age`, `SibSp`, `Parch` i `Fare`. Gracies al coneixement que tenim sobre el dataset em estat capaços de concloure que els valors extrems és poden considerar perfectament valors normals per al dataset i no s'ha procedit a l'eliminació de cap observació.

Hem pogut veure com la proporció entre homes i dones que van sobreviure és bastant diferent. En general és va salvar un tant percent més alt de dones que d'homes. També s'ha pogut comprovar que la relació entre la supervivencia dels nens i els adults també és bastant diferent tenint els nens una tassa de supervivencia més alta. Així doncs podem pensar que el protocol d'ebacuació del vaixell prioritzava a les dones i els nens abans que els homes. Així doncs, un simple agent que respongues sempre Sí a les preguntes on el passatger sigui nen o dona tindrà moltes possiblitats d'encertar, així com repondre No en cas dels homes.

Hem pogut veure gràcies a les proves realitzades sobre el conjunt amb els tests de Kolmogorov-Smirnov i de Shapiro-Wilk, que la variable `Age` no segueix una distribució normal.

2.6 Contribucions al treball

Contribucions	**Signatura**
Investigació prèvia	AMP
Redacció de les respostes	AMP
Desenvolupament codi	AMP

Important: La resolució de la pràctica es pot veure en el següent enllaç https://data-science-env.github.io/Titanic-Data-Analysis/