Pràctica 2 - Tipologia i cicle de vida de les dades

Lídia Bandrés Solé, Guillem Vidal

28/12/2021

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset original s'ha extret del següent enllaç:

Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

```
if (!require('ggplot2')) install.packages('ggplot2')
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
library(ggplot2)
```

```
if (!require('dplyr')) install.packages('dplyr')
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(dplyr)
```

El primer pas serà carregar el dataset per poder-lo analitzar.

```
##   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
Parch
## 1                            Braund, Mr. Owen Harris   male  22     1
```

```
0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
0
## 3                              Heikkinen, Miss. Laina female  26     0
0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
0
## 5                            Allen, Mr. William Henry   male  35     0
0
## 6                                    Moran, Mr. James   male  NA     0
0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

En total tenim 12 variables, anem a analitzar-les una a una:

PassengerId: és un variable int que identifica el número de passatger

Servived: és una variable dicotòmica en format número que indica si el passatger va sobreviure o no a l'enfonsament. 0=No, 1=Yes.

Pclass: és una variable politòmica que ens indica la classe del passatger. 1=primera, 2=segona, 3=tercera

Name: és una variable del tipus char que indica el nom del passatger.

Sex: és una variable dicotòmica en format char que indica el sexe del passatger.

Age: és una variable del tipus int que indica l'edat del passatger.

SibSp: variable del tipus integer que indica el nombre de germans/parella que teníen a bord del Titanic.

Parch: variable del tipus integer que indica el nombre de pares/fills que tenia el passatger a bord del Titanic.

Ticket: variable del tipus char que indica el número del bitllet.

Fare: variable del tipus double que indica el preu del bitllet.

Cabin: variable del tipus char que indica el número de la cabina.

Embarked: variable del tipus char que indica el port del qual va embarcar el passatger. C = Cherbourg, Q = Queenstown, S = Southampton.

Fent una primera ullada al dataset ja podem deduïr quines variables seran menys relevants pel nostre estudi, com per exemple el nom del passatger.

L'enfonsament del Titanic va ser un aconteixement que va sacsejar el món l'any 1912. Des d'aleshores s'han fet estudis, documentals i pel·lícules i s'ha convertit quasi en llegenda.

Ens sembla interessant fer un estudi que analitzi sobretot des del punt de vista dels sobrevivents i deixi una mica la llegenda de banda. Estudiar quines són les variables que van afectar més per determinar qui sobrevivia i qui no, per exemple, veure com el classisme va poder afectar qui es salvava i qui no.

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

Com ja hem comentat abans, l'anàlisi que ens sembla més interessant és fer-ho des del punt de vista dels sobrevivents. No serà necessari reduïr el nombre d'obervacions amb cap tècnica ja que no arriben a les 1000 i per tant són números molt manejables.

Hi ha variables que tindràn més relevància que d'altres, amb algunes es pot fer un estudi de correlació per mera curiositat (per exemple, la relació entre els sobrevivents i el port del qual van embarcar) però hi hauran algunes variables que no s'utilitzaran ni tan sols per fer anàlisis curiosos i per tant, es treuran del dataset. Aquestes són:

PassangerId: és una variable redundant ja que en la creació/visualització de qualsevol dataset la informació que proporciona ja ve donada pel número de fila.

Name: tot i que per la memòria històrica tingui un gran valor, per l'estudi que volem dur a terme no té relevància ja que no ens interessen aquelles variables que tinguin un valor únic per cada passatger.

Ticket: té una relació directa amb *Fare* per tant, tenir les dos és redundant. De les dos, ens quedarem amb *Fare* ja que pot ser útil a l'hora d'analitzar el classisme que va poder influenciar en la supervivència dels passatgers.

Cabin: és una variable de poca relevància. Està buida en molts casos i és una variable que resulta impossible emplenar mitjançant tècniques d'imputació, o la tenim o no la tenim. Com que no ens aporta informació de valor pel nostre estudi, en prescindirem.

Per tant, actualitzarem el dataset amb el filtratge que em mencionat i ens quedarà un dataset amb 8 variables:

```
dades <- dades %>% select(Survived, Pclass, Sex, Age, SibSp, Parch, Fare,
Embarked)
```

Com totes les variables amb les que treballarem procedeixen d'un mateix dataset, no és necessari un procés d'integració. Tot i això, si és necessari més endavant, poden passar per un procés de transformació.

3. Neteja de les dades.

El pas següent consistirà en fer un neteja de les dades. No creiem que un procés de normalització sigui necessari pel nostre dataset. Sí podrà ser interessant comprovar

quina distribució segueixen les variables i discretitzar algunes de les variables per simplificar l'estudi.

Per poder comprovar si les variables segueixen una distribució normal, podem aplicar el test Shapiro. El test shapiro ens indicarà que podem assumir normalitat en la distribució de la variable sempre que el valor de p (p-value) sigui superior a 0.05. Sabem d'entrada que tenim algunes variables que seguiran una distribució binomial (a més a més, en el cas del sexe és conegut que és una distribució binomial amb probabilitat 0.5). Per tant, les úniques variables que seria interessant estudiar-ne la normalitat serien *Age* i *Fare*

```
shapiro.test(dades$Age)

##
##  Shapiro-Wilk normality test
##
## data:  dades$Age
## W = 0.98146, p-value = 7.337e-08

shapiro.test(dades$Fare)

##
##  Shapiro-Wilk normality test
##
## data:  dades$Fare
## W = 0.52189, p-value < 2.2e-16
```

Cap de les dues segueix una distribució normal i aquesta informació ens pot ser d'utilitat més endavant.

La discretització pot ser interessant de cara a la variable *Age* però abans de fer-ho, haurem de tractar els valors buits.

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

Per coneixer els valors buits de les variables, utilitzarem la funció summary.

```
summary(dades)

##     Survived          Pclass          Sex                 Age
##  Min.   :0.0000   Min.   :1.000   Length:891         Min.   : 0.42
##  1st Qu.:0.0000   1st Qu.:2.000   Class :character   1st Qu.:20.12
##  Median :0.0000   Median :3.000   Mode  :character   Median :28.00
##  Mean   :0.3838   Mean   :2.309                      Mean   :29.70
##  3rd Qu.:1.0000   3rd Qu.:3.000                      3rd Qu.:38.00
##  Max.   :1.0000   Max.   :3.000                      Max.   :80.00
##                                                      NA's   :177
##      SibSp           Parch             Fare          Embarked
##  Min.   :0.000   Min.   :0.0000   Min.   :  0.00   Length:891
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:  7.91   Class :character
##  Median :0.000   Median :0.0000   Median : 14.45   Mode  :character
```

```
##  Mean   :0.523   Mean   :0.3816   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   Max.   :512.33
##
```

La funció summary ens dóna molta informació d'interés, per exemple, podem veure que la persona més jove era un bebé d'uns 5 mesos i la persona més gran tenia 80 anys. Pel que fa els valors buits (els zeros no els contem perquè en el nostre cas, els zeros ens donen informació) que és el que ens interessa, veiem que l'únic que en té es la variable *Age*. Per assegurar-nos que no ens perdem cap valor buit que no estigui categoritzat com "NA" farem el següent pas:

```
colSums(is.na(dades))

## Survived   Pclass      Sex      Age    SibSp    Parch     Fare
Embarked
##        0        0        0      177        0        0        0
0

colSums(dades=="")

## Survived   Pclass      Sex      Age    SibSp    Parch     Fare
Embarked
##        0        0        0       NA        0        0        0
2
```

Podem veure que ens dóna quasi la mateixa informació que ja teníem, a excepció d'informació extra sobre dos valors buits de la variable *Embarked*

Pel que fa als valors desconeguts de la variable *Age*, donada la naturalesa d'aquesta, la millor solució serà substituir els valors desconeguts per la mitjana.

```
dades$Age[is.na(dades$Age)] <- mean(dades$Age,na.rm=T)
```

Pel que fa als valors buits de *Embarked*, els substituirem per una U de unknown per poder llegir-ho millor ja que només són dos observacions, no és lo suficientment gran/important per dur a terme tot un estudi de correlació per intentar substituir-ho (i sempre amb el dubte que l'estudi no sigui 100% verídic ja que podria ser que trobéssim una tendència, i just els valors que omplim resulta que no la complien)

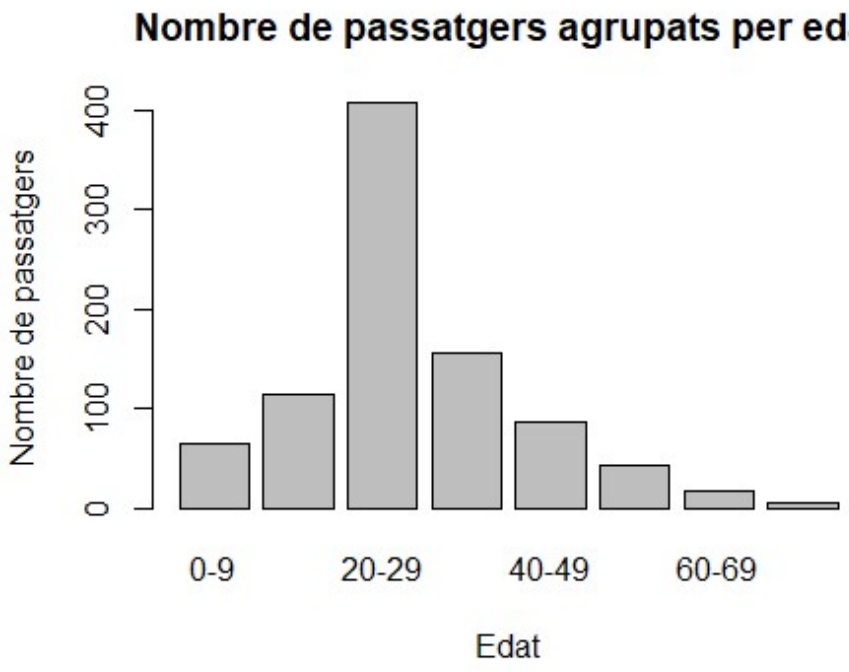
```
dades$Embarked[dades$Embarked==""] <- "U"
```

Un cop hem tractat els valors buits/NA, podem fer la discretització de la variable *Age* que em mencionat anteriorment. Per fer-ho, crearem una nova varia *Age_disc*

```
dades$Age_disc <- cut(dades$Age, breaks = c(0,10,20,30,40,50,60,70,100),
labels = c("0-9", "10-19", "20-29", "30-39","40-49","50-59","60-69","70-
79"))
```

A continuació podem veure com els passatgers s'agrupaven per edat:

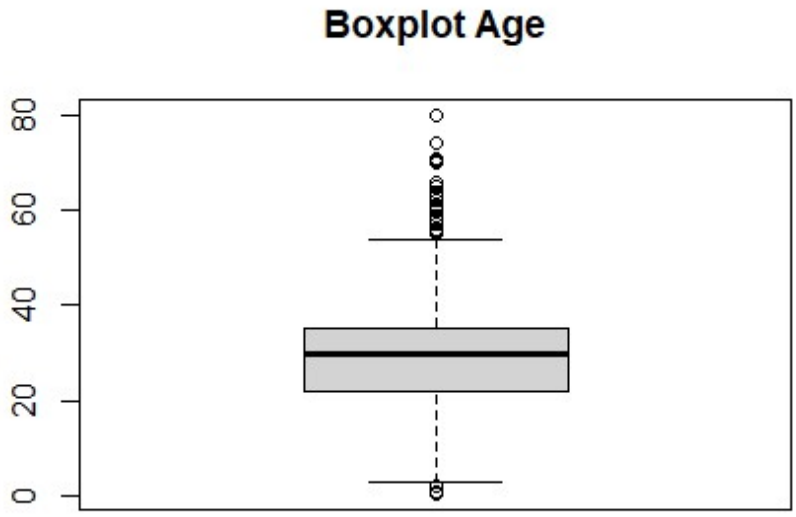
```
plot(dades$Age_disc, main="Nombre de passatgers agrupats per
edat",xlab="Edat", ylab="Nombre de passatgers")
```



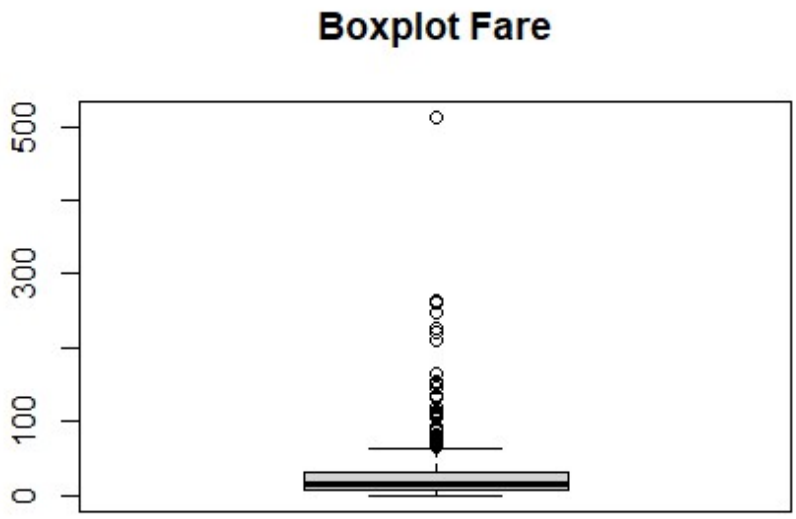
3.2. Identificació i tractament de valors extrems.

Per tal de detectar possibles valors extrems, utilitzarem un diagrama de caixa (boxplot). Les úniques variables on té sentit fer aquest estudi són: *Age* i *Fare*.

```
boxplot_age <- boxplot(dades$Age, main= "Boxplot Age")
```




```
boxplot_fare <- boxplot(dades$Fare, main= "Boxplot Fare")
```



Pel que fa la variable *Age*: ja sabiem d'abans que la mitjana es situava per sota dels 40 anys, el gràfic ens ho il·lustra de forma visual. Veiem que efectivament tenim valors extrems, corresponen per la part baixa als bebés i per la part altra a partir de més o menys 50 anys. Però això no ens fa sospitar que les dades siguin incorrectes ja que és normal que la majoria de passatgers estiguéssin en la trentena, són els que es podien permetre viatjar, molts d'ells amb la família i per això tenim les edats dels nens. Per tant, no s'han de tractar de cap manera ja que són correctes i totes importants per l'anàlisi.

Pel que fa la variable *Fare*: el fet que la mitjana sigui baixa ens demostra que hi havia molts passatgers de tercera classe, és normal que hi hàgi valors extrems per sobre ja que els preus que pagaven els de primera classe podien arribar a ser desorbitats en comparació. També hem de tenir en compte que hi ha uns quants valors 0 (segurament, fan referència a la tripulació). Tot i això, hi ha un valor en concret que crida l'atenció, està per sobre de 500, sol (no en té cap més aprop) i per tant fa sospitar que podria tractar-se d'un error. Si observem les dades del dataset veurem que aquest punt fa referència a tres valors iguals (512.3292) i el següent ja és 263 (casi la mitat).

El tractament que se'ls hi donarà serà el següent: substituirem els valors per la mitjana tots els valors que siguin de primera classe (si fem la mitjana total baixaria massa i no interessa que doni un valor fora de primera classa ja que no seria realista).

```
vector_disc <- vector()

for (i in 1:length(dades$Pclass)){
  if (dades$Pclass[i]==1){
    vector_disc[i] <- dades$Fare[i]
  }
}
vector_disc <- vector_disc[!is.na(vector_disc)] #eliminem valors NA

# Tenim un vector que emmagatzema tots els valors de Fare dels passatgers
de primera classe

mitjana_class1 <- mean(vector_disc)

dades$Fare[dades$Fare==512.3292] <- mitjana_class1
```

Ara ja no tenim el valor extrem.

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

A continuació plantejarem possibles estudis a realitzar:

(1) Relació entre les variables Sex i Survived. Van sobreviure més homes o més dones en total? Si ho comparem amb el nombre de persones de cada sexe que va embarcar, els % de sobreviure segons el sexe són similars?

(2) Relació entre Embarked i Survived. Si hi ha algun port del qual sobreviu més gent, podria ser que en aquell port embarquéssin més persones de primera classe?

(3) Comprovació si l'edat i el preu dels tiquets que tenen els homes i dones de mitjana són els mateixos.

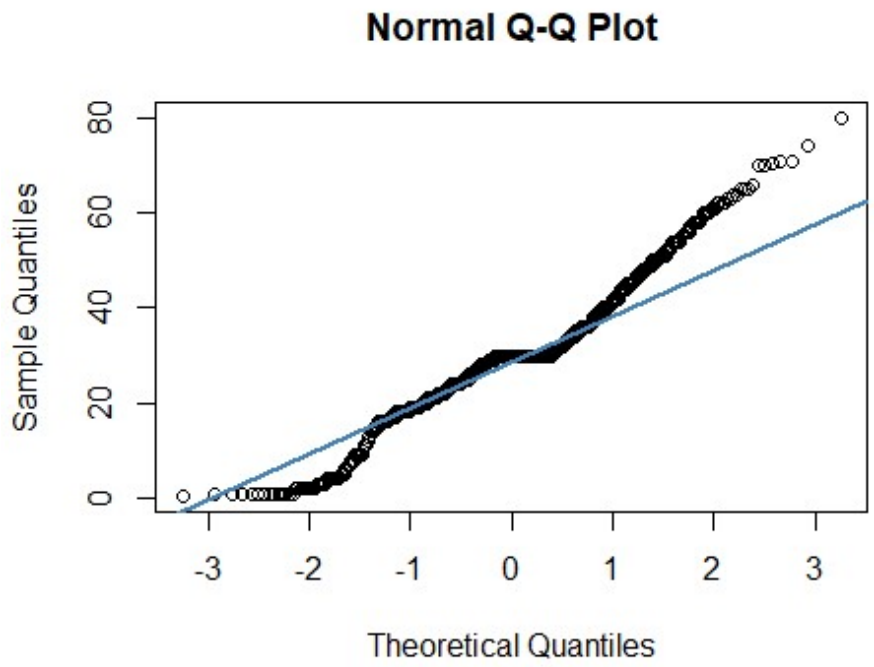
(4) Relació entre Pclass i Survived. Quina va ser la classe que va sobreviure més gent en total? Si ho comparem amb el nombre de persones que hi havia de cada classe, quina classe va tenir el % més alt de sobrevivents?

(5) Relació entre Age i Survived. Es va donar prioritat a les persones grans, o potser als nens? La majoria de passatgers estaven a la trentena, són el grup amb % més alt de sobrevivents?

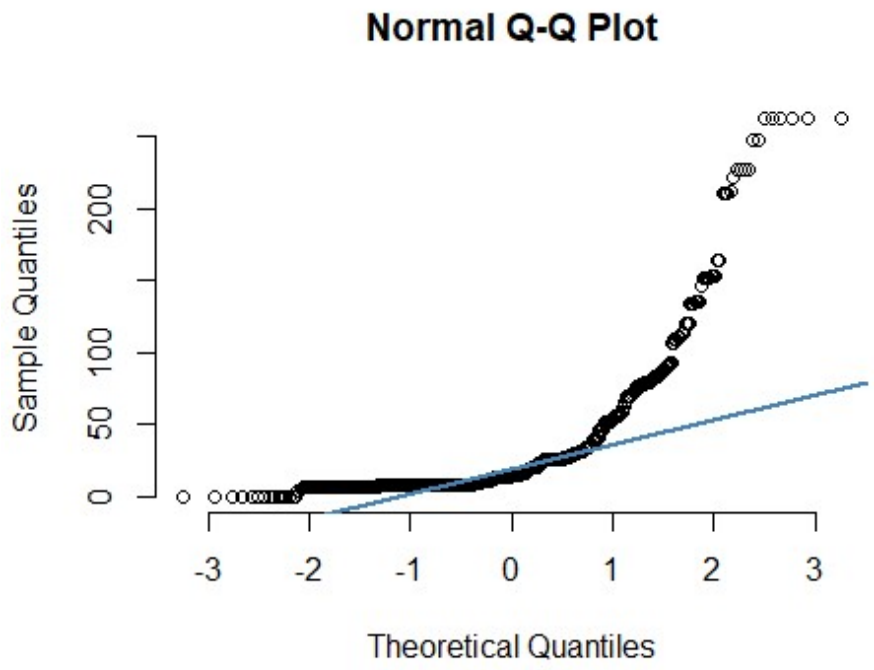
4.2. Comprovació de la normalitat i homogeneïtat de la variància.

En primer lloc fem el contrast de normalitat de les variables numèrques del dataset. Per fer-ho realitzarem el gràfic Q-Q ja que hem ja hem realitzat prèviament el test de Shapiro-Winks (apartat 3)

```
{qqnorm(dades$Age)
qqline(dades$Age, col = "steelblue", lwd = 2)}
```



```
{qqnorm(dades$Fare, pch = 1, frame = FALSE)
qqline(dades$Fare, col = "steelblue", lwd = 2)}
```



Com hem vist anteriorment les dues variables no segueixen una distribució normal.

Tot seguit evaluarem la homogeïtat de la variança. Per fer-ho utilitzarem el test de Leneve ja que no podem asegurar que les variables siguin normals.

```
if (!require('car')) install.packages('car')
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.1.2
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(car)
leveneTest(dades$Age,dades$Fare)
```

```
## Warning in leveneTest.default(dades$Age, dades$Fare): dades$Fare
coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group 247  1.1025 0.1728
##       643
```

Observem que no hi ha diferències significatives entre les variances dels dos grups.

4.3. 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.

En primer lloc mirarem la relació de les persones que van sobreviure segons el sexe. Habitualment es té en compte el fet que les dones i els nens van per davant. D'aquesta forma podrem saber si en el cas del titanic es va complir.

Per fer-ho utilitzarem un analisi univariant amb predictors categòrics:

```
# Utilitzem una regressió logística:
dades$Survived<-as.factor(dades$Survived)
dades$Pclass<-as.factor(dades$Pclass)
model1<-glm(dades$Survived ~ dades$Sex, family = binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = dades$Survived ~ dades$Sex, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6462  -0.6471  -0.6471   0.7725   1.8256
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.0566     0.1290   8.191 2.58e-16 ***
## dades$Sexmale  -2.5137     0.1672 -15.036  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  917.8  on 889  degrees of freedom
## AIC: 921.8
##
## Number of Fisher Scoring iterations: 4
```

Amb una $p < 0.05$ podem dir que l'edat és un bon predictor dels supervivents. To seguit mirem els ODDS:

```
exp(coefficients(model1))
```

```
##   (Intercept) dades$Sexmale
##    2.87654321    0.08096732
```

Els resultats ens indiquen que amb un 95% de confiança les probabilitats de sobreviure sent home són del 8,1% mentre que essent dona són del 91.9%.

Seguit l'exemple del sexe ara mirarem si homes i dones tenen de mitjana la mateixa edat i han pagat el mateix preu d'entrada segons un contrast d'hipòtesis:

```
#En primer lloc dividim el dataset entre homes i dones:
homes = dades[!(dades$Sex == "male"), ]
dones = dades[!(dades$Sex == "female"), ]
```

Suposant normalitat i homocedasticitat definim les hipòtesis:

Hipòtesis nula és

$$H_0: \mu H = \mu D$$

Hipòtesis alternativa és

$$H_1: \mu H \neq \mu D$$

```
#Utilitzem el t.test:

t.test(homes$Age, dones$Age)
```

```
##
##  Welch Two Sample t-test
##
## data:  homes$Age and dones$Age
## 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
```

Amb un p valor de 0'011 podem refusar la hipòtesis nul·la i per tant podem afirmar que hi ha diferències significatives amb les mitjanes de cada sexe.

Realitzem el mateix procediment amb el preu del tiquet:

```
t.test(homes$Fare, dones$Fare)
```

```
##
##  Welch Two Sample t-test
##
## data:  homes$Fare and dones$Fare
## t = 5.9407, df = 449.94, p-value = 5.688e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  12.76575 25.38715
## sample estimates:
## mean of x mean of y
##  43.11620  24.03975
```

Amb un p-valor inferior al 0'05 es trobem en el mateix cas, trobem diferències per sexe sobre el preu del tiquet.

Finalment comprovem la relació entre els anys i la classe dels passatgers.

En primer lloc utilitzem el mètode Chi-quadrat per saber si existeix relació:

```
matriu<-table(dades$Survived, dades$Pclass)
matriu
```

```
##
##       1   2   3
##   0  80  97 372
##   1 136  87 119
```

Definim la hipòtesis nul·la la qual ens diu que no existeix relació entre les dues variables. És a dir, no són dependents.

Calculem l'estadístic de X^2:

```
chisq.test(matriu)
```

```
##
##  Pearson's Chi-squared test
##
## data:  matriu
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

El p-valor ens indica que rebutgem la H0 i, per tant, definim que les variables són depenents.

Ara podem evaluar quina relació o dependència existeix:

```
model2<-glm(dades$Survived ~ dades$Pclass, family = binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = dades$Survived ~ dades$Pclass, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4094  -0.7450  -0.7450   0.9619   1.6836
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.5306     0.1409   3.766 0.000166 ***
## dades$Pclass2  -0.6394     0.2041  -3.133 0.001731 **
```

```
## dades$Pclass3  -1.6704     0.1759  -9.496  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance: 1083.1  on 888  degrees of freedom
## AIC: 1089.1
##
## Number of Fisher Scoring iterations: 4
```

Amb una $p < 0.05$ podem dir que la classe és un bon predictor dels supervivents. To seguit mirem els ODDS:

```
exp(coefficients(model2))
```

```
##   (Intercept) dades$Pclass2 dades$Pclass3
##     1.7000000     0.5275925     0.1881720
```

Podem afirmar que ser de classe 2 o 3 tenies menys probabilitats de sobreviure al tenir coeficients negatius, un 52% en el cas de la classe 2 i un 18% en el cas de la classe 3.

5. Representació dels resultats a partir de taules i gràfiques.

En primer lloc, per obtenir una vista general de les variables establirem un diagrama de correlació:

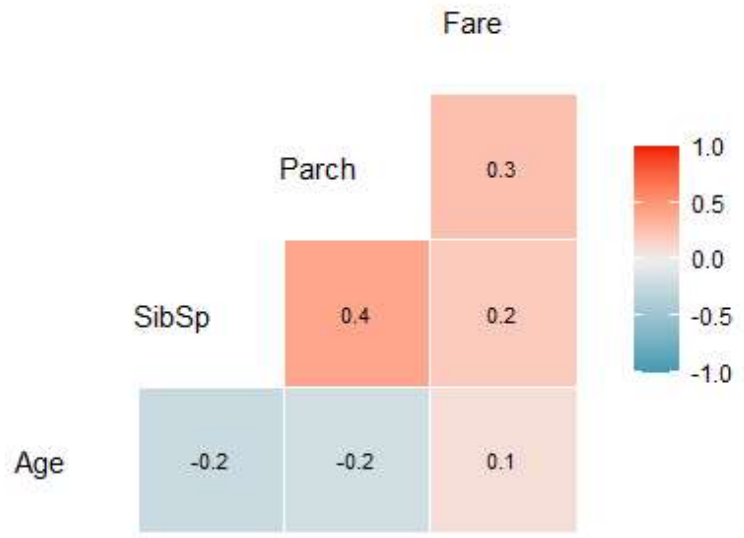
```
library('GGally')
```

```
## Warning: package 'GGally' was built under R version 4.1.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggcorr(dades,label = T,label_size = 2.9,hjust = 1,layout.exp = 1)
```

```
## Warning in ggcorr(dades, label = T, label_size = 2.9, hjust = 1,
layout.exp =
## 1): data in column(s) 'Survived', 'Pclass', 'Sex', 'Embarked',
'Age_disc' are
## not numeric and were ignored
```



En segon lloc anem a visualitzar la relació entre els supervivents i l'edat:

```
if (!require('tidyverse')) install.packages('tidyverse')
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages ---------------------------------------
tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

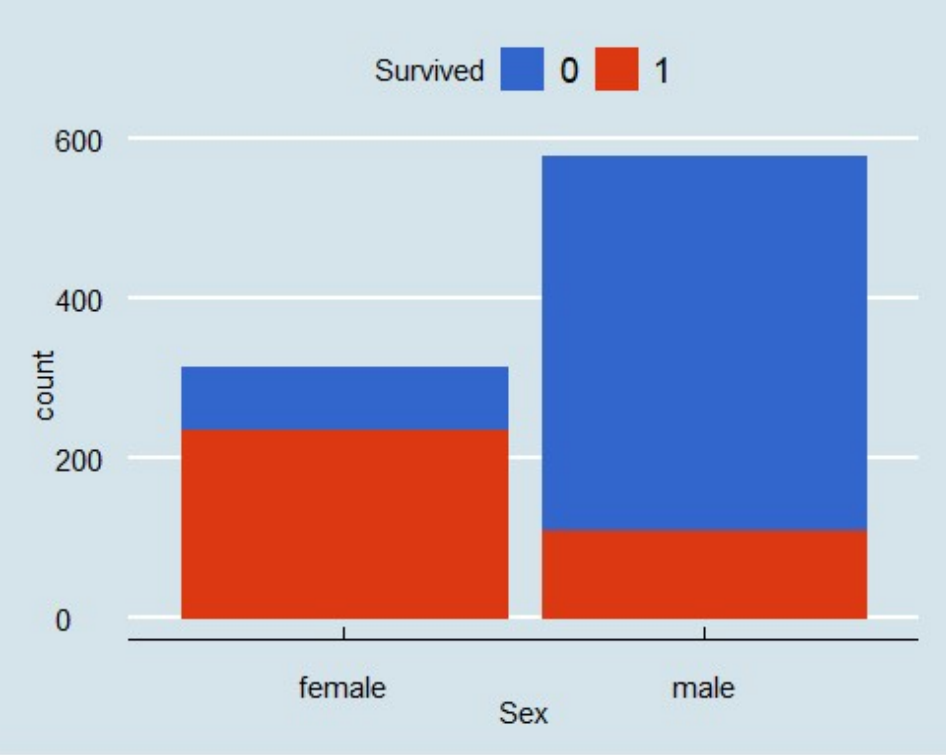
```
## Warning: package 'purrr' was built under R version 4.1.2
```

```
## Warning: package 'stringr' was built under R version 4.1.2
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
```

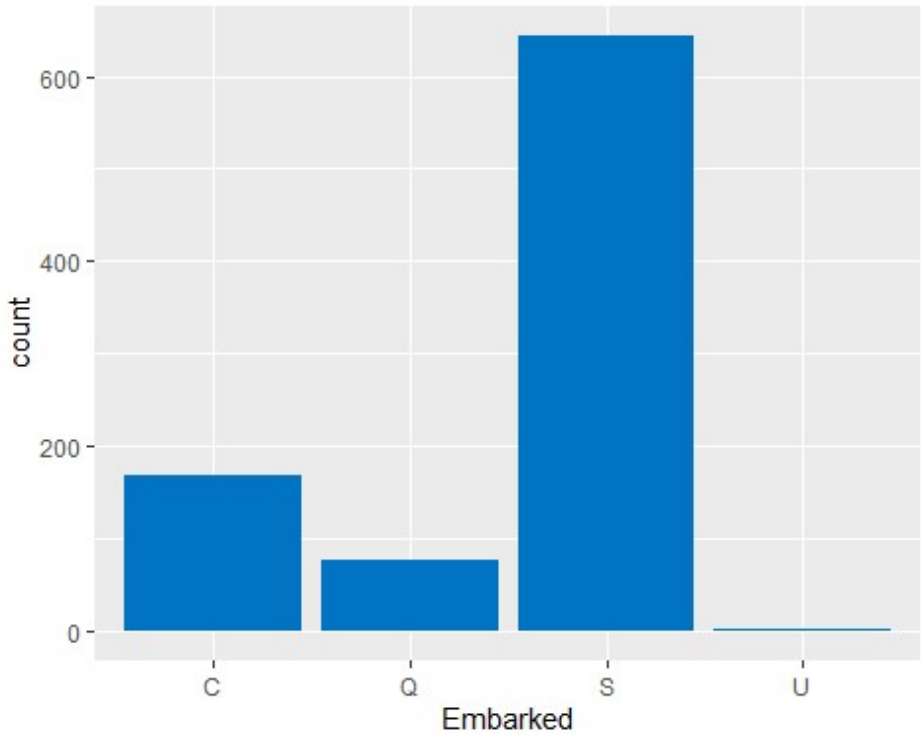
```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x car::recode()   masks dplyr::recode()
## x purrr::some()   masks car::some()

if (!require('ggthemes')) install.packages('ggthemes')

## Loading required package: ggthemes

## Warning: package 'ggthemes' was built under R version 4.1.2

library(tidyverse)
library(ggthemes)
ggplot(dades, aes(x=Sex, fill=Survived, colour=Survived)) + geom_bar() +
ggthemes::theme_economist() + scale_color_gdocs() +
ggthemes::scale_fill_gdocs()
```



Tal com ens ha demostrat el contrast podem saber de forma visual que el percentatge de dones que van sobreviure és molt més alt que el d'homes.

Respecte el segon punt que ens plantegem mirem la freqüència de ports dels passatgers:

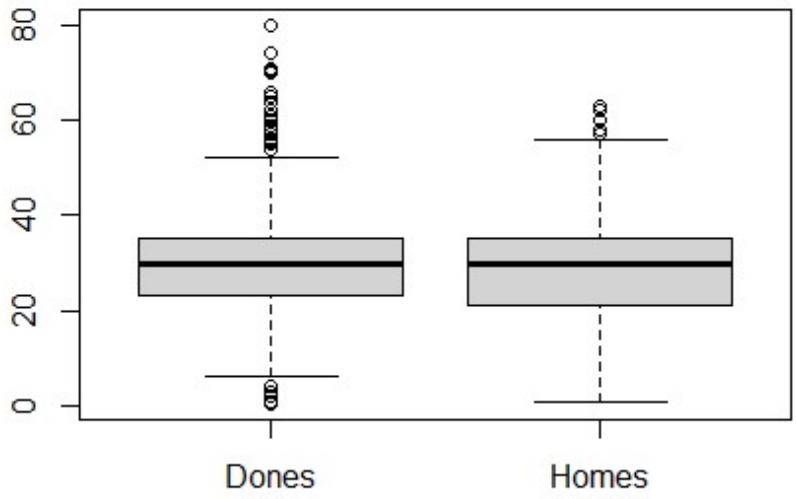
```
ggplot(dades, aes(Embarked)) +
  geom_bar(fill = "#0073C2FF")
```



Observem que la gran majoria de passatgers provenien de Southampton i per tant qualsevol anàlisi hauria tingut poca rellevància.

Tot seguit, il·lustrem la mitjana de d'edat i del preu de l'entrada per tal de trobar-ne differencies:

```
boxplot(dones$Age, homes$Age, names = c("Dones","Homes"))
```



Tot i que el contrast d'hipòtesis ens demostra que les mitjanes d'edat no són iguals, a simple vista no es pot determinar.

```
if(!require('modeest')) install.packages('modeest')
```

```
## Loading required package: modeest
```

```
## Warning: package 'modeest' was built under R version 4.1.2
```

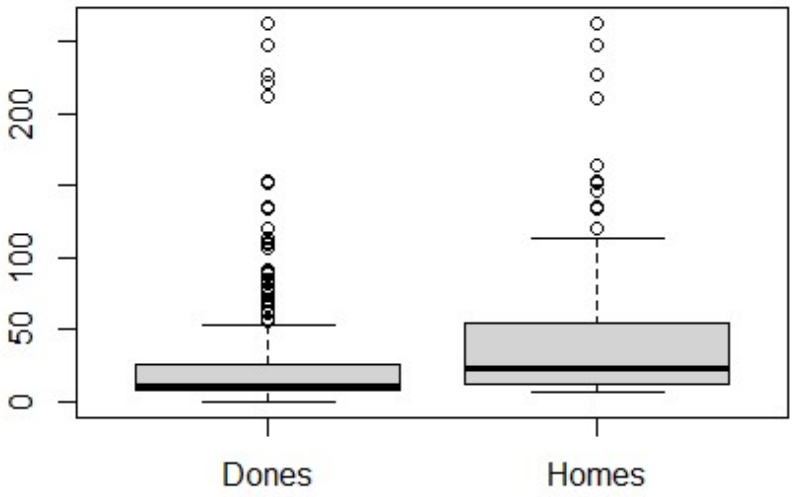
```
## Registered S3 method overwritten by 'rmutil':
##   method         from
##   print.response httr
```

```
library(modeest)
mitH<-round(mean.default(homes$Age), 2)
mitD<-round(mean.default(dones$Age), 2)
Mitjana <- as.numeric(c(mitH, mitD))
Taula <- as.data.frame(rbind(Mitjana))
colnames(Taula)<-c("Homes", "Dones")
Taula
```

```
##         Homes Dones
## Mitjana 28.22 30.51
```

Observem que realment les mitjanes són diferents però visualment no apreciables.

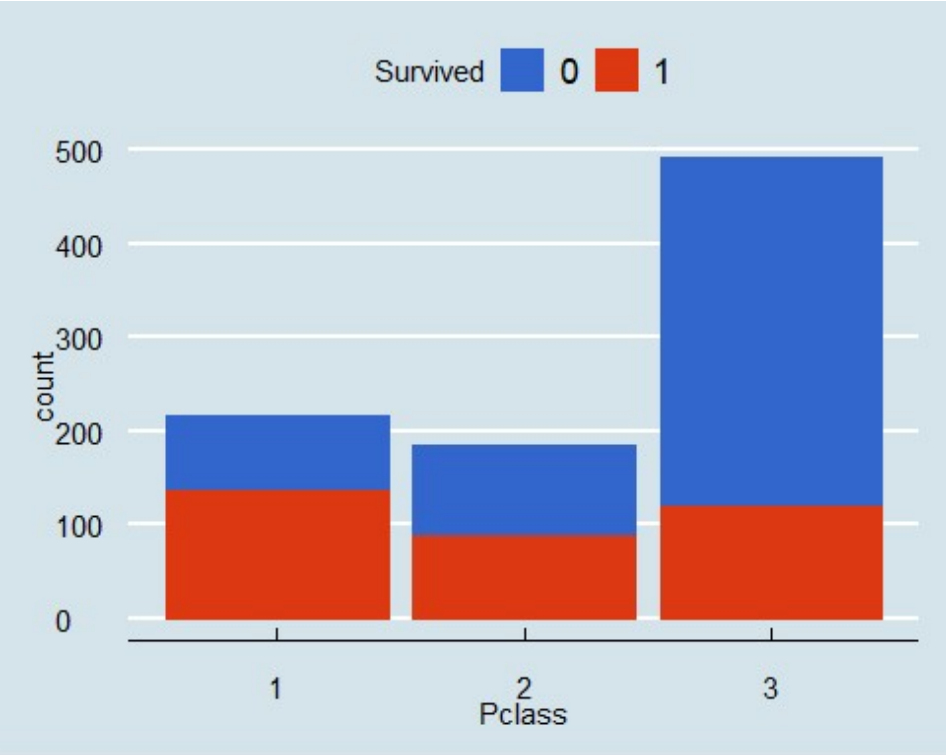
```
boxplot(dones$Fare, homes$Fare, names = c("Dones","Homes"))
```



En aquest cas si que s'aprecia la diferència visualment de les mitjanes. Fet que ens indica que és més destacable la diferència dels preus dels tiquets entre sexes.

Relació entre classe i supervivents:

```
ggplot(dades, aes(x=Pclass, fill=Survived, colour=Survived)) + geom_bar()
+ ggthemes::theme_economist() + scale_color_gdocs() +
ggthemes::scale_fill_gdocs()
```



Observem a simple vista que les tres classes tenen un nombre similar de supervivents però els de classe 3 tenen una proporció molt més reduïda que les classes 1 i 2. Confirmant així les ODDS anteriors.

6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Podem concluïr que els passatgers del titanic partien de condicions molt diferents a l'hora de saber si sobreviurien a un accident. Com hem pogut comprovar les dones van tenir prioritat envers els homes mentre que la primera classe i la segona també van ser prioritzades. Tenim clar doncs que les persones amb més prioritat van ser les dones de primera classe. Les dades ens han permès trobar resposta a gairebé totes les qüestions plantejades i els contrastos ens han permés tenir una certesa estadística dels resultats.

Finalment, exportarem el dataset final a format csv. Aquest codi és útil només per poder extreure el fitxer csv de cara ha ser entregat amb la resta del projecte. Si es vol obtenir el dataset directament del codi, s'haurà d'editar l'enllaç.

```
write.csv(dades,"C:\\Users\\lidia\\Desktop\\MASTER UOC\\M2.951. Tipologia
i cicle de vida de les dades\\Pràctica 2\\Titanic_updated.csv", row.names
= FALSE)
```

Taula de contribucions:

Taula de contribució:

Contribucions	Firma
Investigació prèvia	Guillem Vidal, Lídia Bandrés
Redacció de respostes	Guillem Vidal, Lídia Bandrés
Desenvolupament del codi	Guillem Vidal, Lídia Bandrés