

Tipologia i cicle de vida de  
les dades

# Pràctica 2

---

Marc Trepap

10 de Desembre 2021



## Taula de continguts

<b>Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?</b>	<b>2</b>
<b>Integració i selecció de les dades d'interès a analitzar</b>	<b>3</b>
<b>Neteja de les dades</b>	<b>4</b>
3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?	5
3.2. Identificació i tractament de valors extrems	7
<b>Anàlisi de les dades</b>	<b>9</b>
4.1 Selecció dels grups de dades que es volen analitzar/comparar	9
4.2. Comprovació de la normalitat i homogeneïtat de la variància.	10
4.3. Aplicació de proves estadístiques per comparar els grups de dades.	11
<b>5. Representació dels resultats a partir de taules i gràfiques.</b>	<b>17</b>
<b>6. Resolució del problema.</b>	<b>19</b>

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

El dataset que s'ha triat per a realitzar la pràctica s'ha obtingut de [Kaggle](#) i el CSV es el titanic. Esta format per 12 variables i un total d'observacions de 891, tal i com es pot veure un cop es carrega el csv al programa:

```
library(readr)
titanic <- read_csv("./titanic.csv")
```

```
## Rows: 891 Columns: 12
```

Les descripcions de les variables són:

- PassengerId: variable numerica. Indica el codi de passatger.
  - Exemples: 1 2 3 4 5 6 7 8 9 10 ...
- Survived, variable numerica. Indica si el passatger ha sobreviscut(1) o no(0)
  - Exemples: 0 1 1 1 0 0 0 0 1 1 ...
- Pclass, variable numerica. Indica la tarifa del viatge: 1 = 1st, 2 = 2nd, 3 = 3rd
  - Exemples: 3 1 3 1 3 3 1 3 3 2 ...
- Name, variable alfanumèrica. Indica el nom i cognoms del passatger.
  - Exemples: "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" ...
- Sex, variable alfanumèrica. Indica el sexe del passatge.
  - Exemples: "male" "female" "female" "female" ...
- Age, variable numèrica. Indica l'edat del passatger.
  - Exemples: 22 38 26 35 35 NA 54 2 27 14 ...
- SibSp, variable numèrica. Indica el número de familiars(germans, esposa,marit) te el passatger.
  - Exemples: 1 1 0 1 0 0 0 3 0 1 ...
- Parch, variable numèrica. Indica el numero de fills, pares te el passatger.
  - Exemples: 0 0 0 0 0 0 0 1 2 0 ...
- Ticket, variable numèrica. Indica el numero de bitllet.
  - Exemples: "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
- Fare. variable numerica. Indica el preu del bitllet.

- Exemples: 7.25 71.28 7.92 53.1 8.05 ...
- Cabin, variable numerica. Indica el número de Cabina.
  - Exemples: NA "C85" NA "C123" ...
- Embarked, variable alfanumerica. Indica a quin port ha embarcat el passatger. C = Cherbourg, Q = Queenstown, S = Southampton
  - Exemples: "S" "C" "S" "S" ..

Despres d'analitzar el dataset, es planteja quines persones van sobreviure depenent del tipus de persona i la classe del bitllet, ja que cada tipus de bitllet, s'ubica en una zona o una altra del vaixell també es té en compte el tipus de persona perquè un cop naufragat depen de les variables de persona de si es jove o no, té fills...

Així amb aquest anàlisis es pot extreure quin tipus de persona necessita alguna seguretat extra per poder millorar el percentatge de supervivents en cas de un altre naufragi.

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

Com s'ha comentat al punt anterior les variables que es necessiten són sobre les persones i el tipus de bitllet:

```
> names(titanic_original)
[1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
[6] "Age"          "SibSp"       "Parch"       "Ticket"      "Fare"
[11] "Cabin"        "Embarked"
```

Per tant les variables que es poden eliminar del dataset són: Name, ticket, fare, cabin. També al tenir tota la informació en un mateix dataset no es necessari aplicar cap fusió de altres fonts.

## 3. Neteja de les dades

Es revisa els valors únics que hi ha a les variables del dataset de titanic:

```
head(unique(titanic$PassengerId))
```

```
## [1] 1 2 3 4 5 6
```

```
unique(titanic$Survived)
```

```
## [1] 0 1
```

```
unique(titanic$Pclass)
```

```
## [1] 3 1 2
```

```
unique(titanic$Sex)
```

```
## [1] "male" "female"
```

```
unique(titanic$Age)
```

```
## [1] 22.00 38.00 26.00 35.00 NA 54.00 2.00 27.00 14.00 4.00  
58.00 20.00  
## [13] 39.00 55.00 31.00 34.00 15.00 28.00 8.00 19.00 40.00 66.00  
42.00 21.00
```

```
unique(titanic$SibSp)
```

```
## [1] 1 0 3 4 2 5 8
```

```
unique(titanic$Parch)
```

```
## [1] 0 1 2 5 3 4 6
```

```
unique(titanic$Embarked)
```

```
## [1] "S" "C" "Q" NA
```

A simple vista es pot veure que els valors són consistent i sense cap error de transcripció, lo que es detecta que hi ha alguns casos que dades incompletes que en el següent punt es solucionaran

```
colSums(is.na(titanic))
```

##	PassengerId	Survived	Pclass	Sex	Age	SibSp
##	0	0	0	0	177	0
##	Parch	Embarked	group_age			
##	0	2	177			

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

Com s'ha detectat al punt anterior hi ha les següents variables de Age y Embarked que tenen elements buits:

La variable Age, al tenir informació suficient es pot predir quin valor pot tenir amb el algorisme de KNN en aquest cas es tindrà en compte el sexe del passatger:

```
titanic.knn <- na.omit(titanic)
titanic.knn$Embarked <- as.numeric(as.factor(titanic.knn$Embarked))
```

Es crea una variable `titanic.knn` on es treu els registres NA i per l'algorisme que aplicarem es necessita tenir variables numeriques per això es transforma la variable `Embarked`.

```
predict_age_for_sex_passenger <- function(subset){  
  rows <- sample(1:nrow(subset), 0.8*nrow(subset))  
  train <- subset[rows,]  
  test <- subset[-rows,]  
  model <- knn(train, test ,train$Age)  
  
  result <- table(model)  
  return(as.integer(names(result[result==max(result)])) )  
}
```

Es crea una funció on es prepara les dades que se li passa per parametre i s'executa el proces KNN per a obtenir el valor.

```
#Male  
titanic.knn.male <- titanic.knn[titanic.knn$Sex == "male",c(1,2,3,5,6,7)]  
age.male <- mean(predict_age_for_sex_passenger(titanic.knn.male))  
  
#Female  
titanic.knn.female <- titanic.knn[titanic.knn$Sex == "female",c(1,2,3,5,6,7)]  
age.female <- mean(predict_age_for_sex_passenger(titanic.knn.female))
```

Es guarda el resultat de la funció amb una variable i es realitza la mitjana per si ens retorna dos valors ja que si fos el cas que hi hagués dos resultat amb el valor maxim.

Un cop es te identificat les edats per a cada sexe, ens toca informar-ho al nostre dataset a les observacions que no tenen valor:

```
table(is.na(titanic$Age))
```

```
##  
## FALSE TRUE  
##    714  177
```

S'informa les observacions que no tenen edat:

```
titanic[titanic$Sex == "male" & is.na(titanic$Age),]$Age <- age.male
titanic[titanic$Sex == "female" & is.na(titanic$Age),]$Age <- age.female
table(is.na(titanic$Age))
```

```
##
## FALSE
## 891
```

La següent variable que s'ha de treballar es **Embarked**, en aquest cas és buscare quin es el valor que mes es repeteix:

```
table(titanic$Embarked)
```

```
##
## C  Q  S
## 168 77 644
```

En aquest cas es pot veure que el valor es **S** amb 644 observacions, per tant, es prepara el codi per a informar del camp de les observacions:

```
titanic[is.na(titanic$Embarked),]$Embarked <- "S"
```

Es crea una variable categorica de Embarked:

```
titanic["Embarked_cat"] <- 3
titanic[titanic$Embarked == "S",]$Embarked_cat <- 1
titanic[titanic$Embarked == "C",]$Embarked_cat <- 2
```

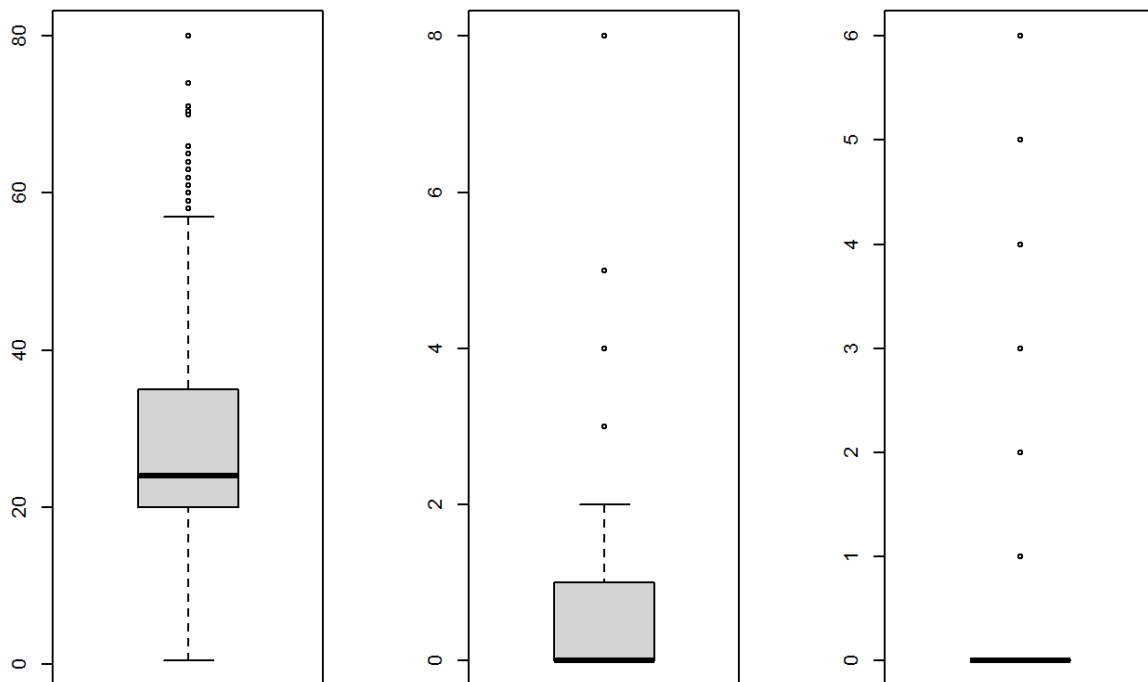
## 3.2. Identificació i tractament de valors extrems

Per a la identificació de valors extrems s'utiliza per a variables numerica continues i per tant es comprova per a edat, SibSp i Parch:



```
par(mfrow=c(1,3))
outliers.age <- boxplot(titanic$Age)
outliers.sib <- boxplot(titanic$SibSp)
outliers.par <- boxplot(titanic$Parch)
```

El resultat son 3 gràfiques on es mostren si hi ha outliers, es pot veure que hi ha outliers amb aquest dataset:



Per a revisar els valors outliers i poder determinar quin tractament s'ha de aplicar es realitza:

```
outliers.age$out
```

```
## [1] 58.0 66.0 65.0 59.0 71.0 70.5 61.0 58.0 59.0 62.0 58.0 63.0 65.0 61.0 60.0  
## [16] 64.0 65.0 63.0 58.0 71.0 64.0 62.0 62.0 60.0 61.0 80.0 58.0 70.0 60.0 60.0  
## [31] 70.0 62.0 74.0
```

```
outliers.sib$out
```

```
## [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3 5 4 3  
## [39] 4 8 4 3 4 8 4 8
```

```
outliers.par$out
```

```
## [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1 2 1  
## [38] 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1 1 2 1 2  
## [75] 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2 2 3 4 1 2 1  
## [112] 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1 2 1 1 2 5 2 1 1  
## [149] 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 3 2 1 1 1 1 2 1 2 3 1 2 1 2 2  
## [186] 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 1 1 1 1 3 2 1 1 1 1 5 2
```

Amb els valors outliers anteriors pot haver casos que les persones siguin molt grans o molt petites igual que poder viatgin soles o amb família i que per tant aquests registres no s'haurien de omitir ja que es tindran en compte els passatger amb aquestes característiques.

## 4. Anàlisi de les dades

### 4.1 Selecció dels grups de dades que es volen analitzar/comparar

Els grups de dades que es vol seleccionar son les observacions on els passatgers han sobreviscut dels que no.

```
titanic.survived <- titanic[titanic$Survived == 1,]  
titanic.no.survived <- titanic[titanic$Survived == 0,]
```

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

Per a la comprovació de normalitat es realitzarà amb el test de Kolmogorov-Smirnov dels dos grups creats anteriorment sobre l'edat:

```
y <- pnorm  
ks.test(titanic.survived$Age, y, mean(titanic.survived$Age), sd(titanic.survived$Age))
```

```
## Warning in ks.test(titanic.survived$Age, y, mean(titanic.survived$Age), : ties  
## should not be present for the Kolmogorov-Smirnov test
```

```
##  
## One-sample Kolmogorov-Smirnov test  
##  
## data: titanic.survived$Age  
## D = 0.093327, p-value = 0.005172  
## alternative hypothesis: two-sided
```

Es comprova l'edat del dataset que no han sobreviscut:

```
ks.test(titanic.no.survived$Age, y, mean(titanic.no.survived$Age), sd(titanic.no.survived$Age))
```

```
## Warning in ks.test(titanic.no.survived$Age, y, mean(titanic.no.survived$Age), :  
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##  
## One-sample Kolmogorov-Smirnov test  
##  
## data: titanic.no.survived$Age  
## D = 0.15476, p-value = 7.585e-12  
## alternative hypothesis: two-sided
```

Com es pot veure els resultats, els dos grups no mantenen una distribució normal perquè el pvalor és inferior al 0.05

### 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.*

Es realitzaran les proves de correlacions, contrast d'hipòtesis i regressió lineal per a poder identificar quines variables tenen relació o no amb la supervivència del passatger:

S'analitza les variables amb els passatgers supervivents que depenen de la variable survived:

```
library(corrplot)
```

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

```
## corrplot 0.92 loaded
```

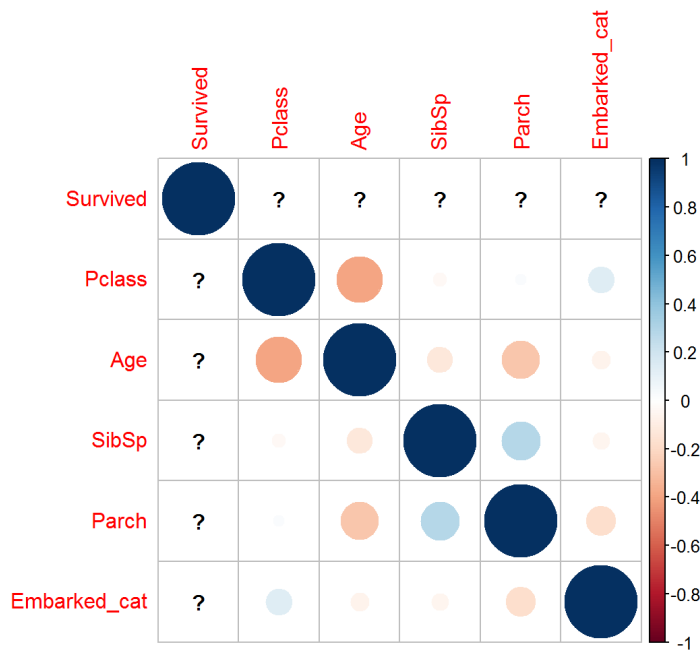
```
c <- cor(titanic.survived[,c("Survived", "Pclass", "Age", "SibSp", "Parch", "Embarked_cat")])
```

Es general les correlacions on es mostra a la següent taula:

c

```
##           Survived      Pclass      Age      SibSp      Parch
## Survived         1         NA         NA         NA         NA
## Pclass           NA  1.00000000 -0.39641210 -0.03329960  0.02158367
## Age             NA -0.39641210  1.00000000 -0.12488929 -0.27573559
## SibSp           NA -0.03329960 -0.12488929  1.00000000  0.28249844
## Parch           NA  0.02158367 -0.27573559  0.28249844  1.00000000
## Embarked_cat    NA  0.13358784 -0.06458582 -0.05385516 -0.17002727
##
##           Embarked_cat
## Survived              NA
## Pclass              0.13358784
## Age                -0.06458582
## SibSp              -0.05385516
## Parch              -0.17002727
## Embarked_cat      1.00000000
```

També es pot generar visualment:



Com es pot veure al gràfic de correlacions la columna Survived no té cap relació forta entre les variables, el que si es pot identificar la relació de la edat del passatger i el tipus de bitllet.

Es prova amb un altre model a veure si ens dona més informació sobre les relacions de Survived.

```
summary(lm(Survived ~ Pclass, data = titanic))
```

```
##
## Call:
## lm(formula = Survived ~ Pclass, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.6416 -0.2476 -0.2476  0.3584  0.7524
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.83863    0.04510   18.60  <2e-16 ***
## Pclass      -0.19700    0.01837  -10.72  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4581 on 889 degrees of freedom
## Multiple R-squared:  0.1146, Adjusted R-squared:  0.1136
## F-statistic: 115 on 1 and 889 DF, p-value: < 2.2e-16
```

```
summary(lm(Survived ~ Sex, data = titanic))
```

```
##
## Call:
## lm(formula = Survived ~ Sex, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7420 -0.1889 -0.1889  0.2580  0.8111
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.74204    0.02307   32.17  <2e-16 ***
## Sexmale     -0.55313    0.02866  -19.30  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4087 on 889 degrees of freedom
## Multiple R-squared:  0.2952, Adjusted R-squared:  0.2944
## F-statistic: 372.4 on 1 and 889 DF, p-value: < 2.2e-16
```

```
summary(lm(Survived ~ Age, data = titanic))
```

```
##
## Call:
## lm(formula = Survived ~ Age, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4222 -0.3933 -0.3687  0.6095  0.6891
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.423569    0.038135  11.107  <2e-16 ***
## Age         -0.001408    0.001222  -1.152   0.249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4865 on 889 degrees of freedom
## Multiple R-squared:  0.001492, Adjusted R-squared:  0.0003684
## F-statistic: 1.328 on 1 and 889 DF, p-value: 0.2495
```

```
summary(lm(Survived ~ SibSp, data = titanic))
```

```
##
## Call:
## lm(formula = Survived ~ SibSp, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3920 -0.3920 -0.3764  0.6080  0.6704
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.39199    0.01804   21.726  <2e-16 ***
## SibSp       -0.01559    0.01479   -1.054    0.292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4866 on 889 degrees of freedom
## Multiple R-squared:  0.001248, Adjusted R-squared:  0.0001242
## F-statistic: 1.111 on 1 and 889 DF, p-value: 0.2922
```

```
summary(lm(Survived ~ Parch , data = titanic))
```

```
##
## Call:
## lm(formula = Survived ~ Parch, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.6607 -0.3650 -0.3650  0.6350  0.6350
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.36503    0.01799   20.294  <2e-16 ***
## Parch        0.04928    0.02018    2.442   0.0148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4852 on 889 degrees of freedom
## Multiple R-squared:  0.006663, Adjusted R-squared:  0.005546
## F-statistic: 5.963 on 1 and 889 DF, p-value: 0.0148
```

```
summary(lm(Survived ~ Embarked_cat, data = titanic))
```

```
##
## Call:
## lm(formula = Survived ~ Embarked_cat, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5178 -0.3543 -0.3543  0.5639  0.6457
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.27253    0.03835   7.107 2.44e-12 ***
## Embarked_cat   0.08176    0.02553   3.203  0.00141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4841 on 889 degrees of freedom
## Multiple R-squared:  0.01141,    Adjusted R-squared:  0.0103
## F-statistic: 10.26 on 1 and 889 DF,  p-value: 0.001408
```

Com es pot veure els pvalors que mes influeixen per sobreviure es el sexe i el tipus de classe del bitllet, en canvi, l'edat dels passatgers no te gairebe cap relació amb sobreviure.

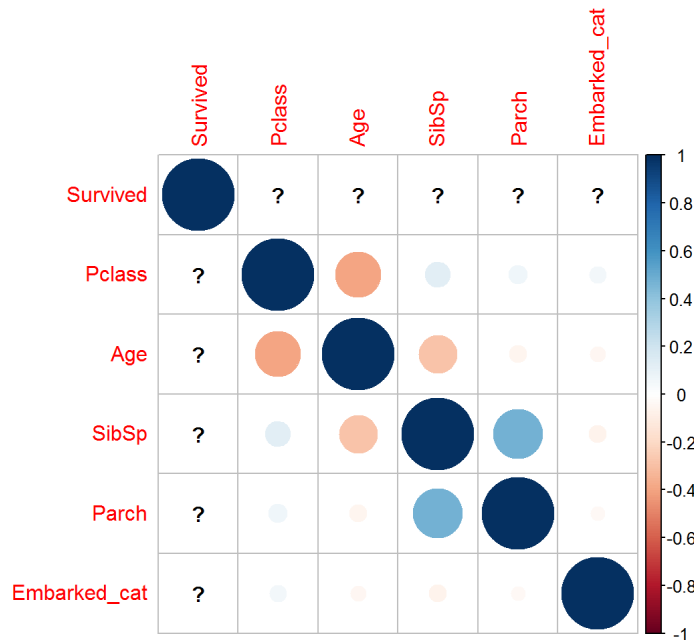
Ara es comprova entre els dos grups si els que sobreviuen son els mes joves o no, per això es àprova amb la hipotesi nul·la i alternativa, generant avanç una grafica de correlacions:



```
cn <- cor(titanic.no.survived[,c("Survived", "Pclass", "Age", "SibSp", "Parch", "Embarked_cat")])
```

```
## Warning in cor(titanic.no.survived[, c("Survived", "Pclass", "Age", "SibSp", :  
## the standard deviation is zero
```

```
corrplot(cn)
```



Les hipotesis que es vol estudiar son les següents:

$H_0: \mu_s \leq \mu_k$

$H_1: \mu_s > \mu_k$

**On:**  $\mu_s$  es la mitjana d'edat dels supervivents,  $\mu_k$  es la mitjana d'edat dels no supervivents.

Com s'ha analitzat al punt anterior la variable de Age no te una distribució normal, donat que cada grup es major que 30 observacions es pot conciderar que el t-test sera suficientment robust:

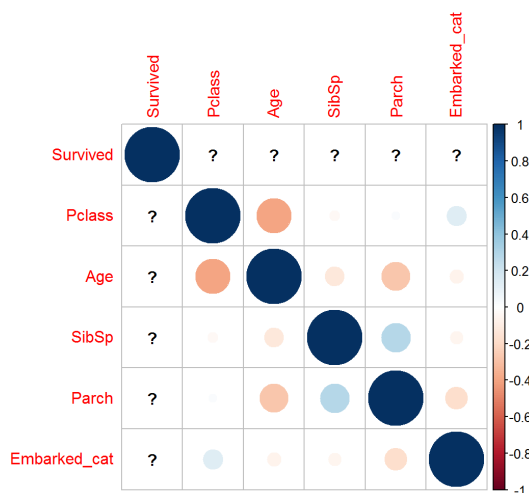
```
t.test(titanic.survived$Age,titanic.no.survived$Age,alternative="greater", v
```

```
##  
## Welch Two Sample t-test  
##  
## data: titanic.survived$Age and titanic.no.survived$Age  
## t = -1.1346, df = 686.88, p-value = 0.8715  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -2.597344 Inf  
## sample estimates:  
## mean of x mean of y  
## 27.56629 28.62568
```

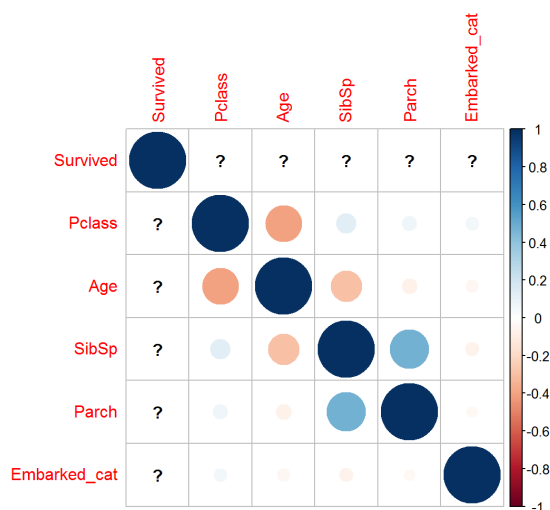
Com es pot veure el pvalue es superior al 0.05 i per tant s'accepta la hipotesis nul·la amb el resultat que la mitjana de edat dels supervivents es la més jove o de la mateixa edat.

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

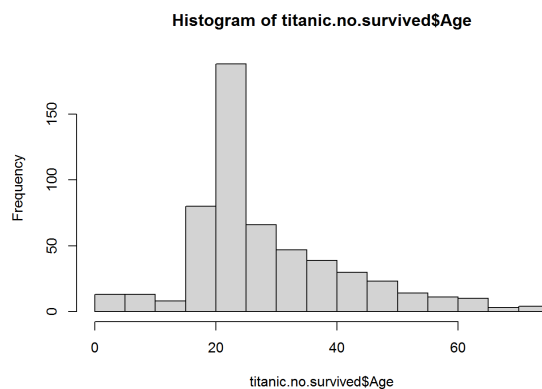
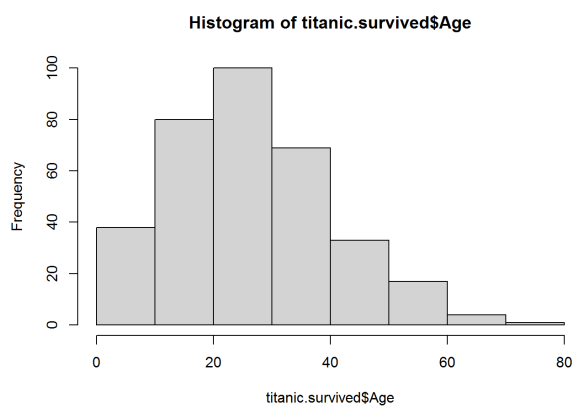
Com s'ha vist al punt anterior sobre el grup de supervivents hi ha les següents correlacions:

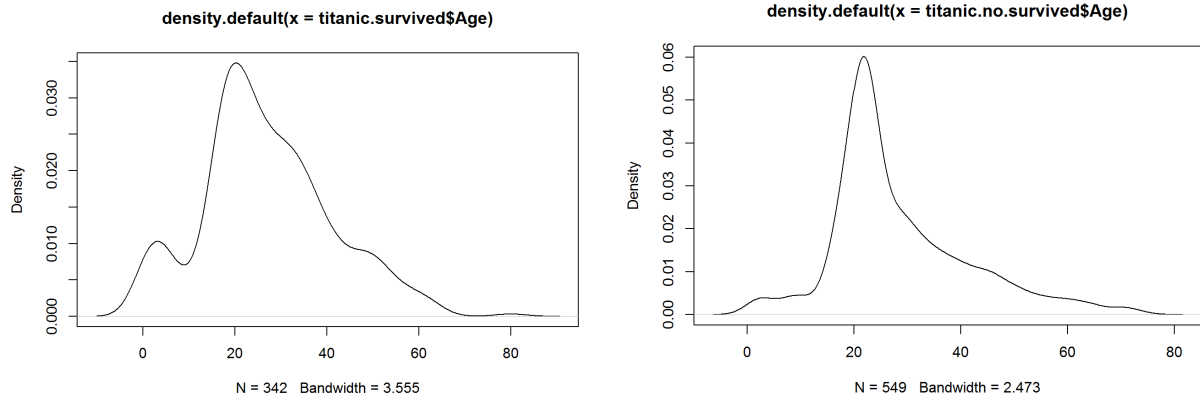


Si es mira al grup que no ha sobreviscut:

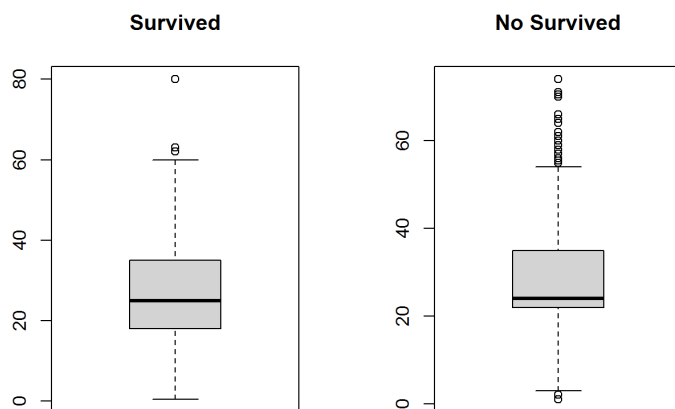


El resultat es molt semblant entre els dos grups, També per a comprovar si l'edat te una distribució normal tant pels supervivents com no, es pot realitzar per un histograma o un gràfic de densitat:






Com es pot veure la primera es una distribució binomial i la segona no té una distribució normal. Es pot analitzar si la hipotesis és correcta a través del gràfic del boxplot:



Com es pot veure els supervivents tenen una edat més alta que els que van morir, amb el resultat anterior de la hipotesis s'afirmava que els supervivents tenien menys edat que els joves.

## 6. Resolució del problema.

*A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?*



Inicialment s'ha proposat poder identificar quins passatgers han sobreviscut amb les variables de persona i del tipus de bitllet, les conclusions que s'ha pogut estudiar, es que no depen d'aquestes variables si sobreviure o no ja que a la gràfica de correlacions no s'ha pogut identificar cap relació, s'ha aplicat una regressió lineal simple sobre la variable sobreviure i s'ha identificat que hi ha una relació amb les variables d'edat i el tipus de bitllet, poder es degut a qui te mes edat es pot permetre un tipus de bitllet més car i tenir un lloc millor posicionat en el vaixell, en canvi, no te cap influencia l'edat del passatger.

Si s'estudia els passatgers que no han sobreviscut, s'arriba a la mateixa conclusió. Per a cada grup s'identifica que les persones més joves tenen una probabilitat de sobreviure.

Així que acabar costa molt poder identificar les persones que poden sobreviure, amb referencia a la hipotesis les persones joves tenen una possibilitat de viure i les persones que hagin comprat algun tipus de bitllet mes elevat.

