Tipologia i cicle de vida de les dades: PRAC2

Autor: Josep Tormo Costa i Oriol Bardés Robles - https://github.com/jotorcos/titanic-ml

Desembre 2020

Contents


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

URL de github amb el codi de la pràctica

https://github.com/jotorcos/titanic-ml

- Josep Tormo Costa: jotorcos
- Oriol Bardés Robles: obr-uoc

Descripció del dataset.

Dataset sobre els passatgers del Titanic està integrat pels conjunts d'entrenament (891 registres) i de prova (418 registres) disponibles a Kaggle:

https://www.kaggle.com/c/titanic/data

Els 12 camps usats en el dataset són els següents:

• PassengerId: identificador numèric de cada passatger embarcat • Survived: Indica si el passatger va sobreviure o va morir (1 = Survived, 0 = Died) • Pclass: Indica el tipus de ticket (1 = Primera classe, 2 = Segona classe, 3 = Tercera classe) • Name: Nom complet del passatger • Sex: Gènere del passatger

(Male/Female) • Age: Edat del passatger • SibSp: Número de germans/cònjuges a bord entre els passatgers • Parch: Número de pares/fills a bord entre els passatgers • Ticket: Número de ticket del passatger • Fare: Preu del ticket del passatger • Cabin: Número de la cabina assignada al passatger • Embarked: Port on el passatger ha embarcat (C = Cherbourg, Q = Queenstown, S = Southampton)

Integració, selecció i neteja de les dades.

```
# Carreguem els dos fitxers de train i test
train <- read.csv('../data/train.csv', stringsAsFactors = FALSE)
test <- read.csv('../data/test.csv',stringsAsFactors = FALSE)

# Creem una nova columna identificadora per saber si la fila és de train o test.
train$Set <- "Train"
test$Set <- "Test"

# Unim els dos jocs de dades en un només
totalData <- bind_rows(train,test)
files=dim(train)[1]

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

```
## 'data.frame':    1309 obs. of  13 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)"
##  $ 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" ...
##  $ Set        : chr  "Train" "Train" "Train" "Train" ...
```

Treballem els atributs amb valors buits.

```
# Estadístiques de valors buits
print('Valors buits:')
```

```
## [1] "Valors buits:"
```

```
print('NA')
```

```
## [1] "NA"
```

```
print(colSums(is.na(totalData)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
##         Set
##           0
```

```
print("")
```

```
## [1] ""
```

```
print(colSums(totalData==""))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
##         Set
##           0
```

Tenim valors buits en Survived, Age, Fare, Cabin i Embarked.

En l'atribut Survived és normal que hi haja valors buits ja que hem unit el dataset de train amb test, i aquest últim no especifica la clase.

En l'atribut Age sí que falten molts valors, imputarem el valor mitjà.

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

En Fare només falta 1, que resulta ser un passatger de 3a classe, així que podriem imputar el valor mitjà de la taxa dels passatgers de 3a classe.

```
# Prenem la mitjana de la 3a classe per a valors buits de la variable "Fare"
totalData$Fare[is.na(totalData$Fare)] <-
  mean(totalData$Fare[totalData$Pclass == 3], na.rm = T)
```

En Embarked només falten 2, així que els imputarem per la moda.

```
table(totalData$Embarked)
```

```
##
##       C   Q   S
##   2 270 123 914
```

Prenem valor "S" per als valors buits de la variable "Embarked", que és la més comuna.

```
totalData$Embarked[totalData$Embarked == ""] = "S"
```

En Cabin també falten però és un atribut que no ens interessa, perquè no considerem que aporte informació relevant.

```
totalData$Cabin <- NULL
head(totalData)
```

```
##   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
## 1                             Braund, Mr. Owen Harris   male 22.00000     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00000     1
## 3                              Heikkinen, Miss. Laina female 26.00000     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00000     1
## 5                            Allen, Mr. William Henry   male 35.00000     0
## 6                                    Moran, Mr. James   male 29.88114     0
##   Parch           Ticket    Fare Embarked   Set
## 1     0        A/5 21171  7.2500        S Train
## 2     0         PC 17599 71.2833        C Train
## 3     0 STON/O2. 3101282  7.9250        S Train
## 4     0           113803 53.1000        S Train
## 5     0           373450  8.0500        S Train
## 6     0           330877  8.4583        Q Train
```

Discretitzem quan té sentit i en funció de cada variable.

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

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##        1309           3           3        1307           2          99
##       SibSp       Parch      Ticket        Fare    Embarked         Set
##           7           8         929         282           3           2
```

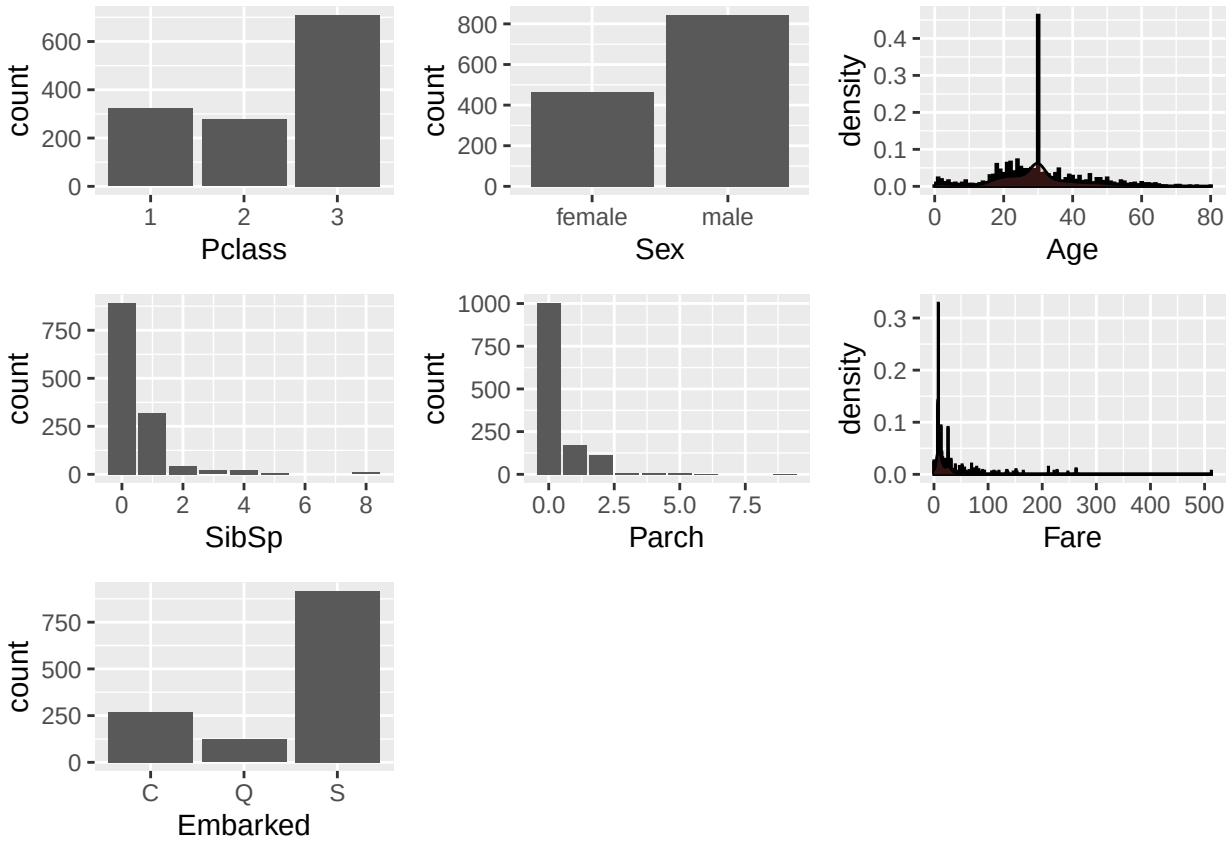
```
# Discretitzem les variables amb poques classes
cols<-c("Survived","Pclass","Sex","Embarked")
for (i in cols){
  totalData[,i] <- as.factor(totalData[,i])
}

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

```
## 'data.frame':    1309 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)"
##  $ 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 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ Set        : chr  "Train" "Train" "Train" "Train" ...
```

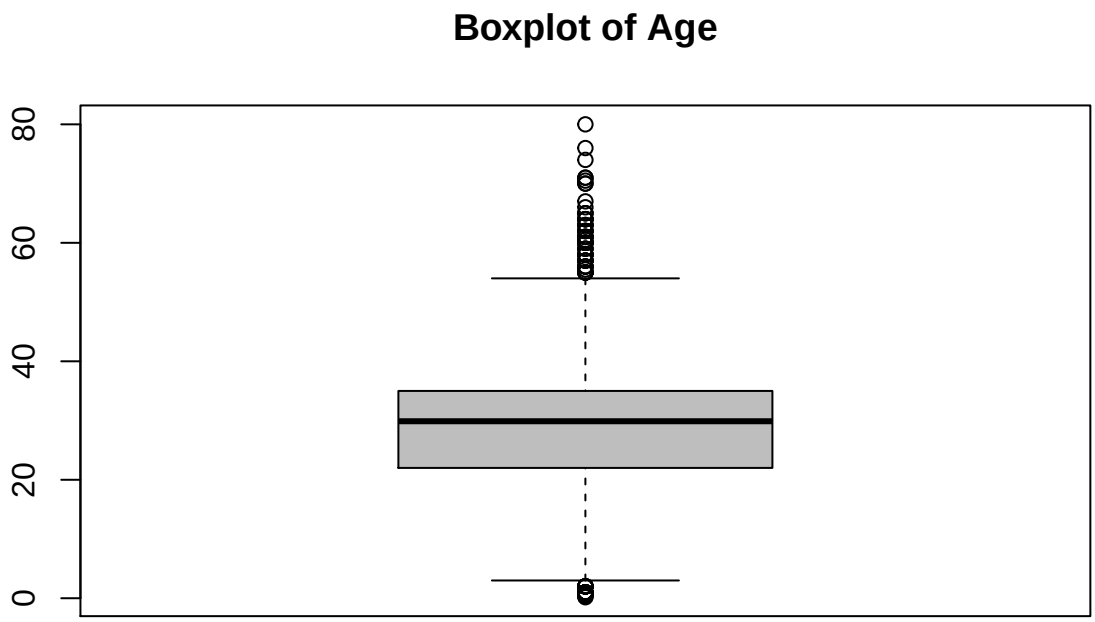
Visualitzem la distribució de les variables:

```
plotDistribution = function (my_data, my_column_name) {
  ggplot(my_data, aes_string(x = my_column_name)) +
    geom_histogram(
      aes(y = ..density..),
      binwidth = .5,
      colour = "black",
      fill = "white"
    ) +
    geom_density(alpha = .2, fill = "#FF6666")
}

p1 <- ggplot(totalData, aes(x = Pclass)) + geom_bar()
p2 <- ggplot(totalData, aes(x = Sex)) + geom_bar()
p3 <- plotDistribution(totalData, 'Age')
p4 <- ggplot(totalData, aes(x = SibSp)) + geom_bar()
p5 <- ggplot(totalData, aes(x = Parch)) + geom_bar()
p6 <- plotDistribution(totalData, 'Fare')
p7 <- ggplot(totalData, aes(x = Embarked)) + geom_bar()

grid.arrange(p1, p2, p3, p4, p5, p6, p7, nrow = 3)
```

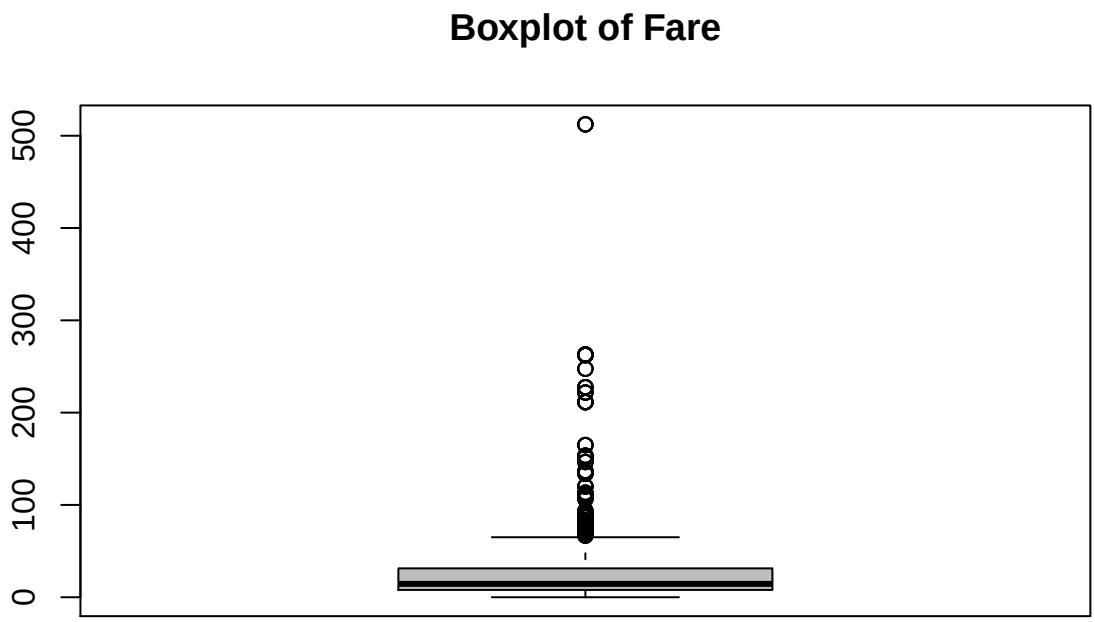


Les variables categòriques no tenen outliers, així que representem els diagrames de caixes per a veure si hi ha valors atípics en les variables continues:

```
p1_box <- boxplot(totalData$Age, main="Boxplot of Age", col="gray")
```




```
p2_box <- boxplot(totalData$Fare, main="Boxplot of Fare", col="gray")
```



Com veiem tenim outliers tant en la variable Age com en Fare.

Per una part, dividirem Age en dos categories: Child i Adult.

```
totalData$AgeGroup[totalData$Age < 18] <- "Child"
totalData$AgeGroup[totalData$Age >= 18] <- "Adult"
totalData$AgeGroup <- as.factor(totalData$AgeGroup)
```

I dividirem Fare en tres categories: Low, Medium, High.

```
totalData$FareGroup <- cut(totalData$Fare, breaks = 3, labels = c("Low", "Medium", "High"))
totalData$FareGroup <- as.factor(totalData$FareGroup)
table(totalData$FareGroup)
```

```
##
##    Low Medium   High
##   1271     34      4
```

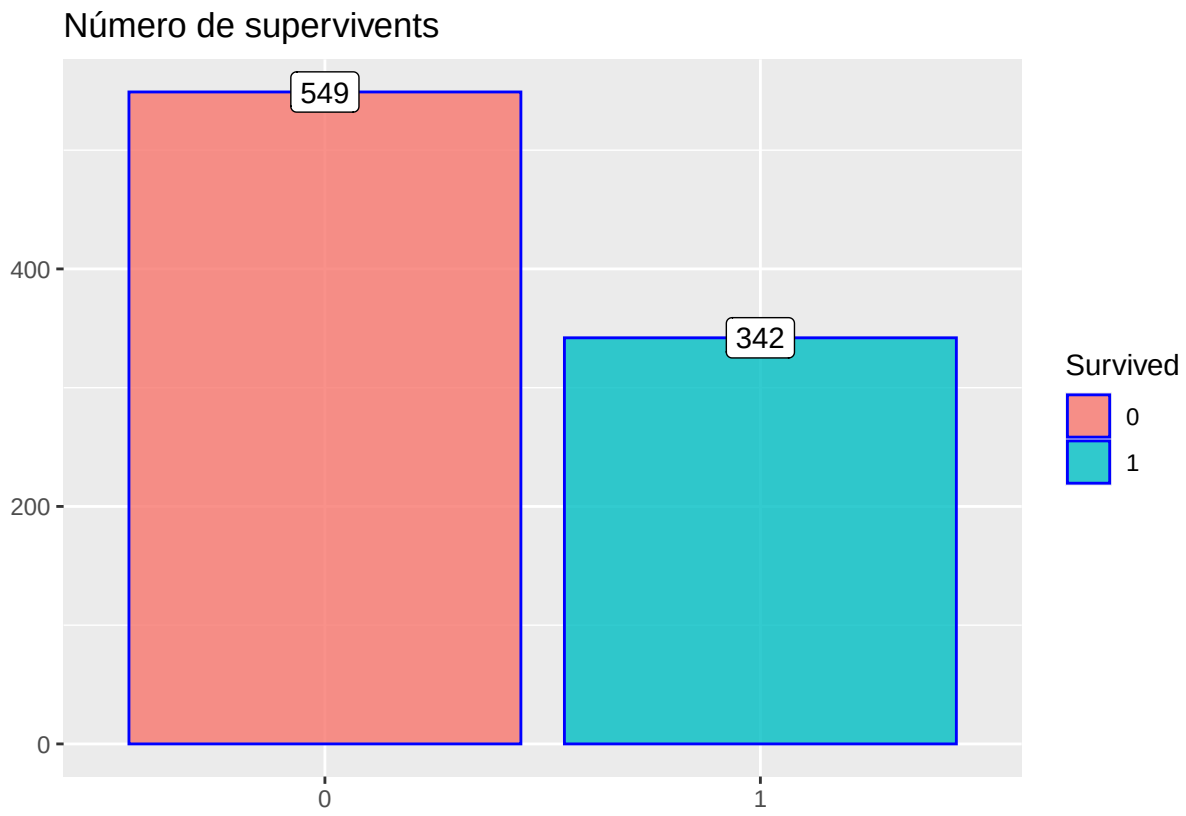
Anàlisi de dades

Ara hem de tornar a dividir el dataset amb el joc de train i test original.

```
train <- totalData[totalData$Set == "Train", ]
test <- totalData[totalData$Set == "Test", ]
```

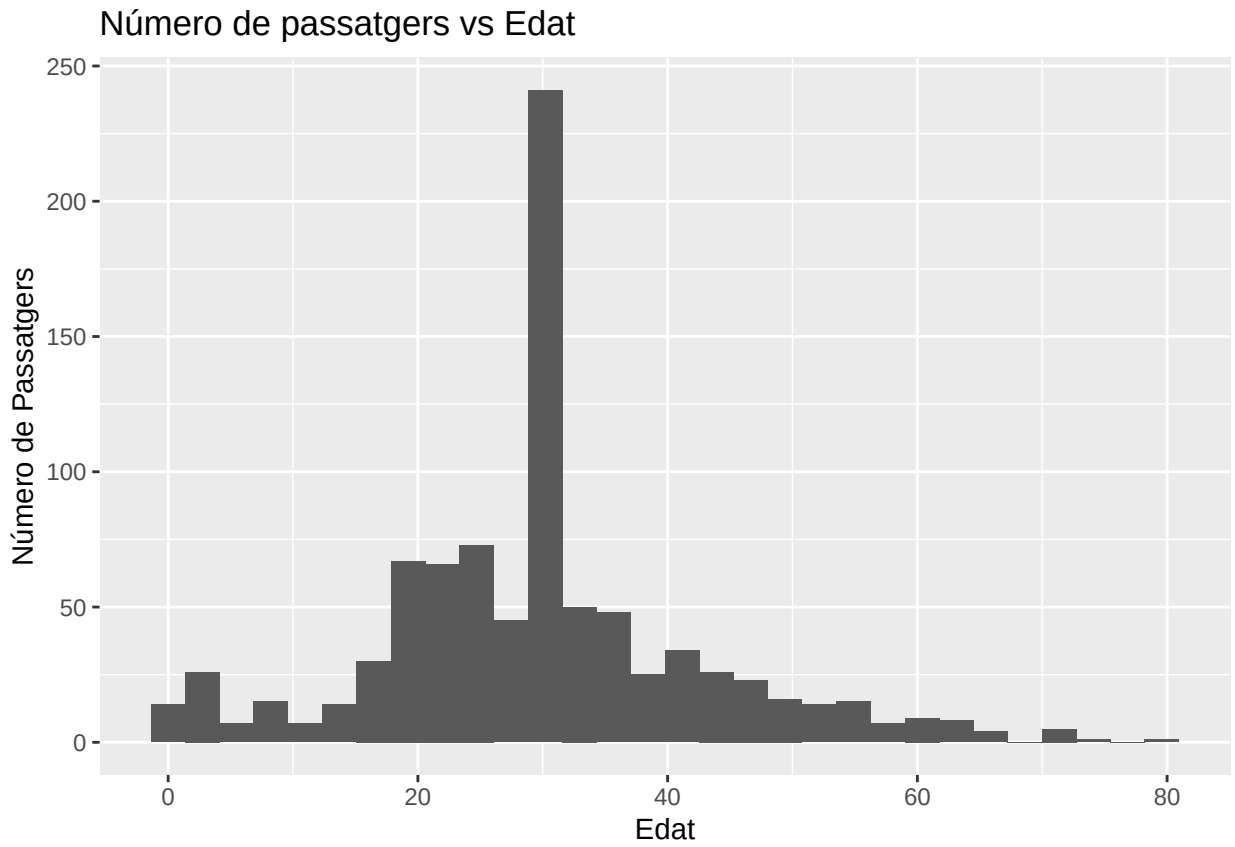
Primerament grafiquem en un histograma la distribució de passatgers supervivents i dels que van morir.

```
ggplot(train, aes(x = Survived)) +
    geom_bar(stat = "count", aes(fill = Survived), col = "blue", alpha = 0.8) +
    labs(x = "", y = "", title = "Número de supervivents") +
    geom_label(stat = "count", aes(label = ..count..))
```



D'aquesta primera anàlisi en resulta que la majoria de passatgers no van sobreviure (549 survivors vs 342 baixes)

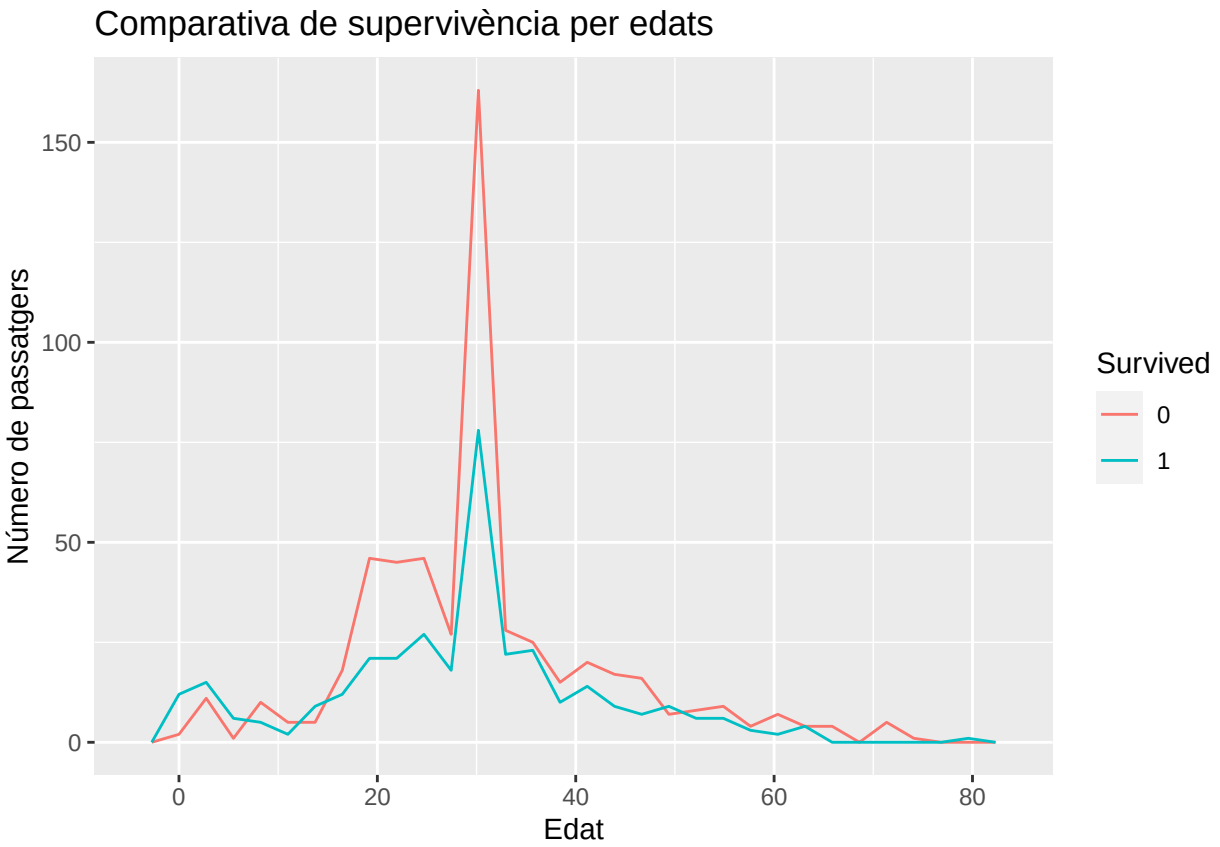
```
ggplot(train, aes(Age)) +
    geom_histogram() +
    labs(x = "Edat", y = "Número de Passatgers", title = "Número de passatgers vs Edat")
```



De l'histograma anterior es desprén que la gran majoria dels passatgers tenien menys de 40 anys.Per tant, podem considerar que el passatge era eminentment jove.

Quina relació hi va haver entre l'edat dels passatgers i la supervivència?

```
ggplot(train) +
  geom_freqpoly(mapping = aes(x = Age, color=Survived)) +
  labs(x = "Edat", y = "Número de passatgers", title = "Comparativa de supervivència per edats")
```

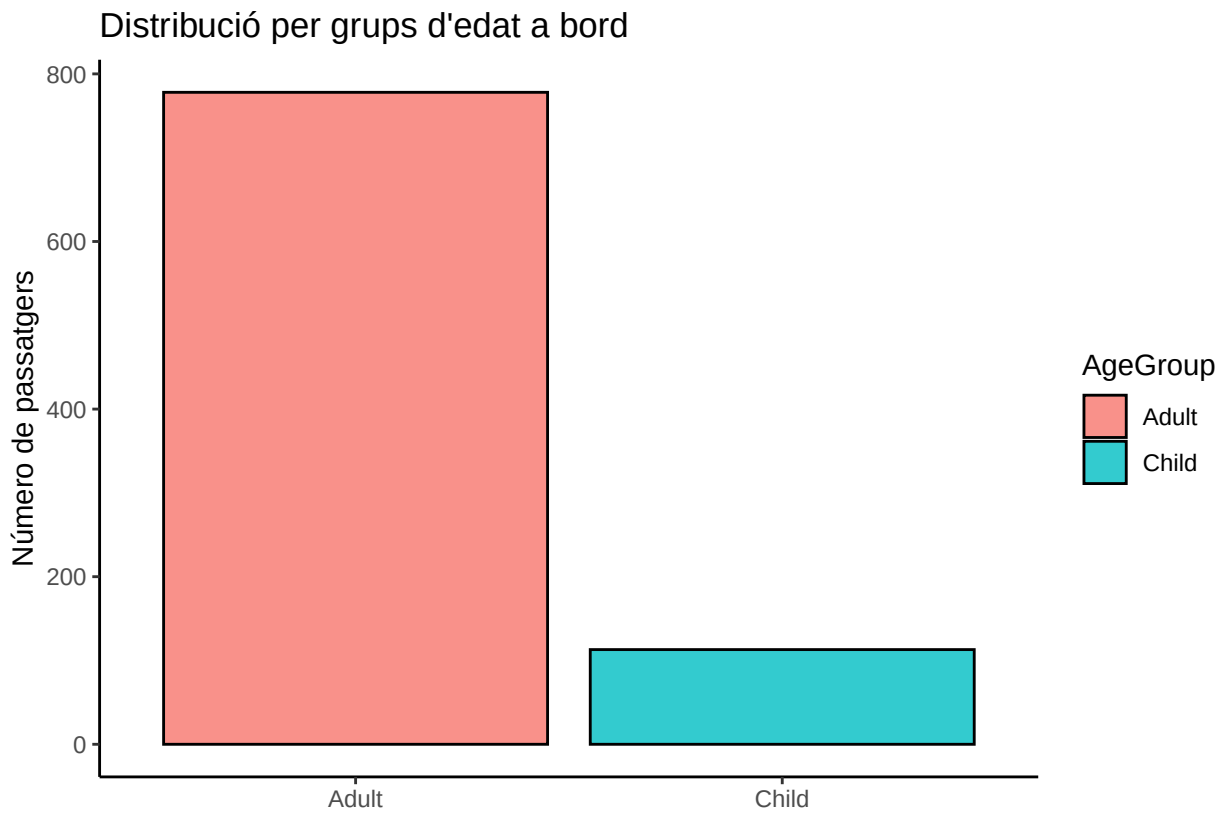


En la gràfica s'aprecia que la supervivència sembla reduïr-se amb l'edat del passtager, cosa que a priori podia intuir-se. Per seguir analitzant aquesta relació, és interessant conèixer la distribució d'adults i menors entre els passatgers:

```
agegroupCount <- train %>%
    group_by(AgeGroup) %>%
    count(AgeGroup) %>%
    select(AgeGroup, Passengers = n)
pandoc.table(agegroupCount)
```

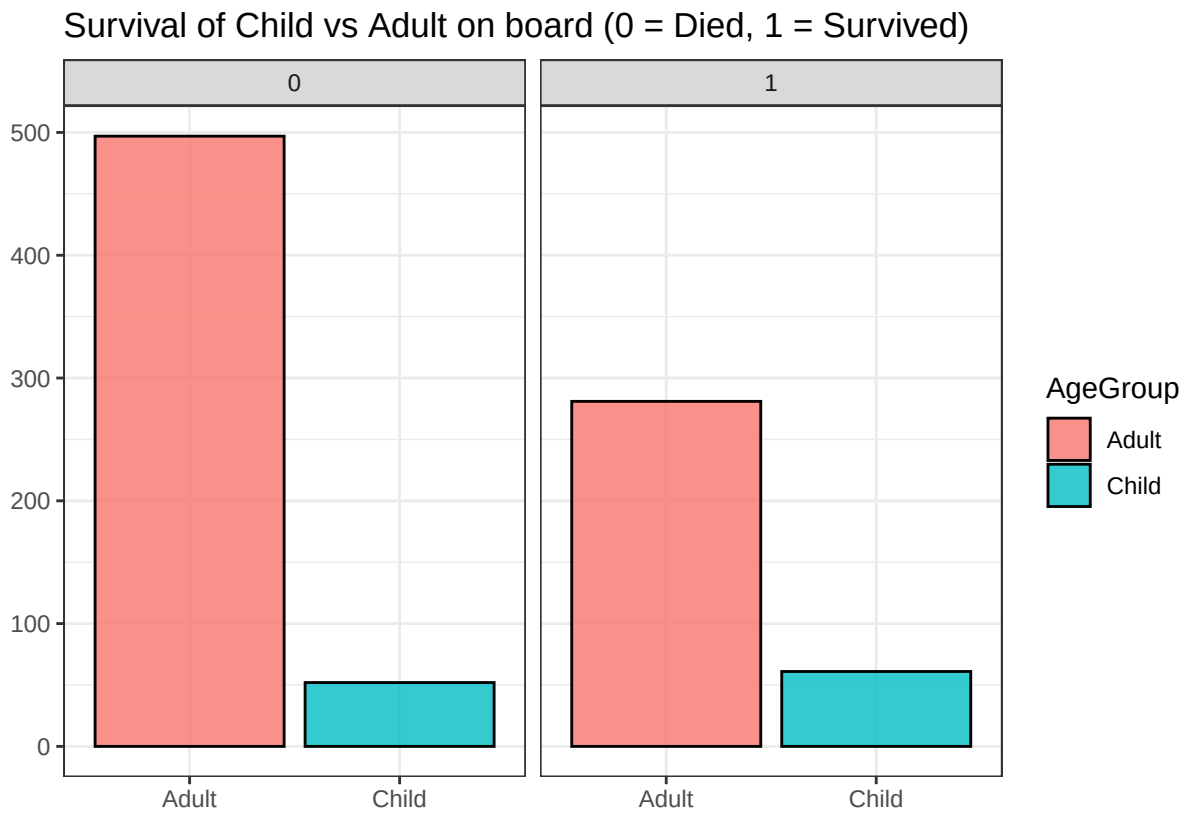
```
##
## ----------------------
##  AgeGroup   Passengers
## ---------- ------------
##   Adult         778
##
##   Child         113
## ----------------------
```

```
ggplot(agegroupCount, aes(x = AgeGroup, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = AgeGroup), col = "black", alpha = 0.8) +
    labs(x = "", y = "Número de passatgers", title = "Distribució per grups d'edat a bord") +
    theme_classic()
```



Es veu clarament que només una petitat petitat part dels passatgers tenien menys de 18 anys.

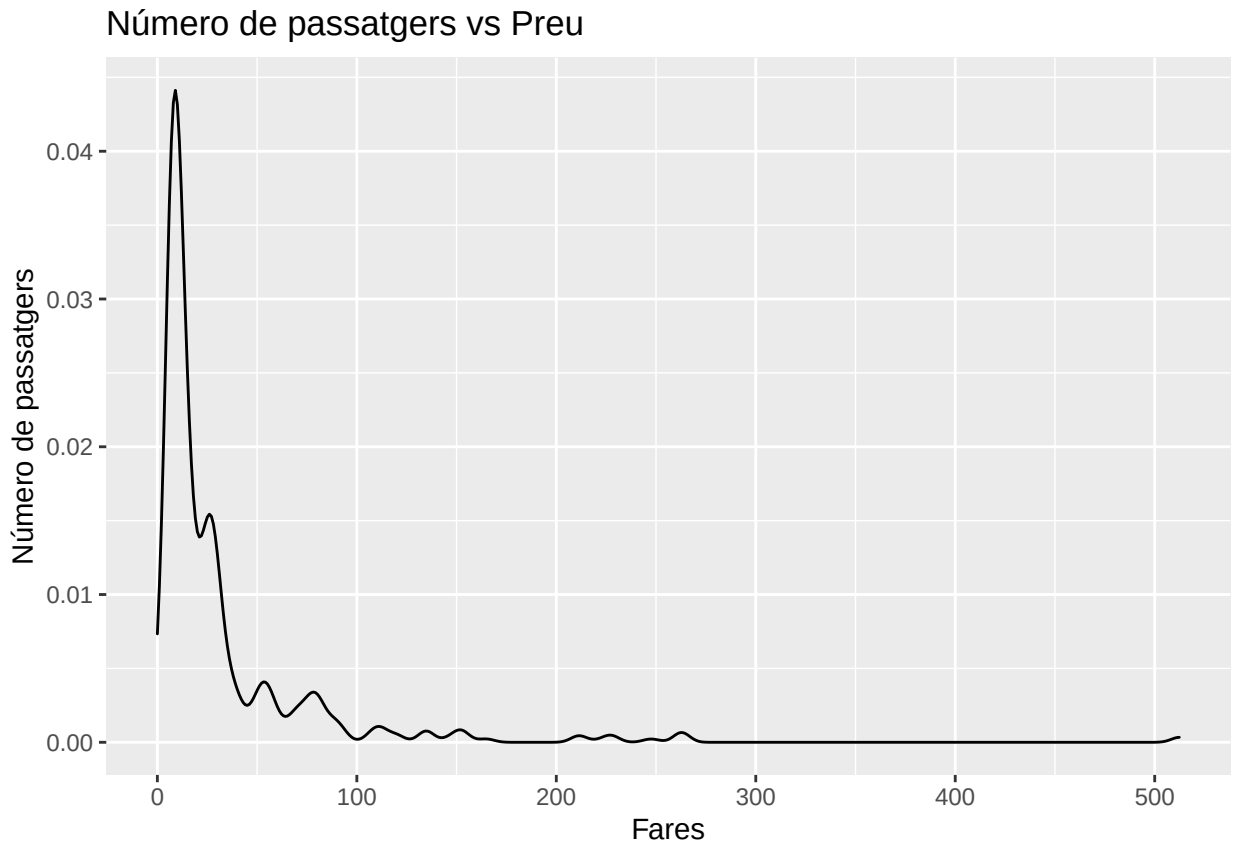
```
agegroupSurvived <- train %>%
    group_by(AgeGroup) %>%
    count(Survived) %>%
    select(AgeGroup, Survived, Passengers = n)

ggplot(agegroupSurvived, aes(x = AgeGroup, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = AgeGroup), col = "black", alpha = 0.8) +
    labs(x = "", y = "", title = "Survival of Child vs Adult on board (0 = Died, 1 = Survived)") +
    facet_wrap(~Survived) +
    theme_bw()
```



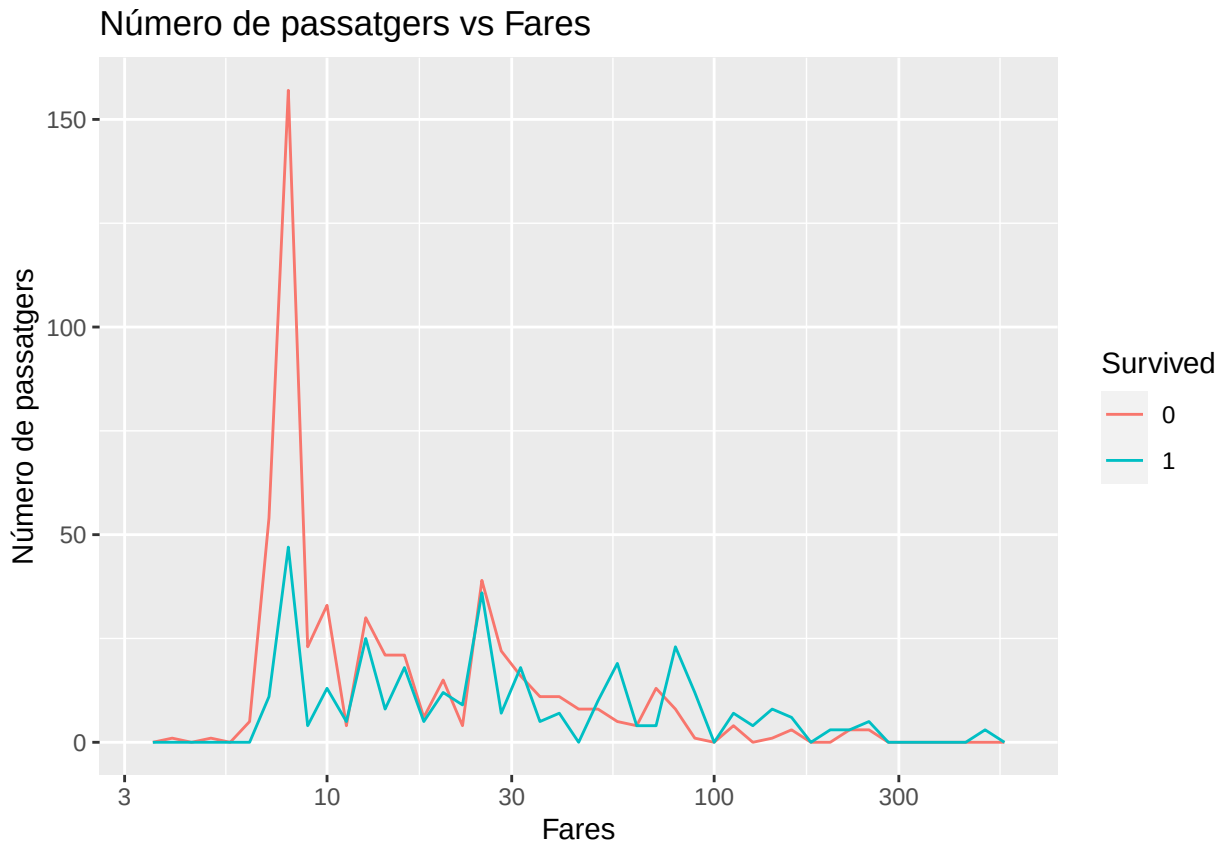
Es veu en el diagrama anterior que la probabilitat de supervivència dels menors d'edat (més del 50%) és sensiblement superior a la dels adults (menys del 40%)

Un altre factor d'anàlisi és la relació entre el preu del bitllet i la supervivència. Repetim l'anàlisi anterior substituint l'atribut Age per Fare:

```
ggplot(train, aes(Fare)) +
    geom_density() +
    labs(x = "Fares", y = "Número de passatgers", title = "Número de passatgers vs Preu")
```




```
ggplot(train) +
  geom_freqpoly(mapping = aes(x = Fare, color = Survived), binwidth = 0.05) +
  scale_x_log10() +
  labs(x = "Fares", y = "Número de passatgers", title = "Número de passatgers vs Fares")
```

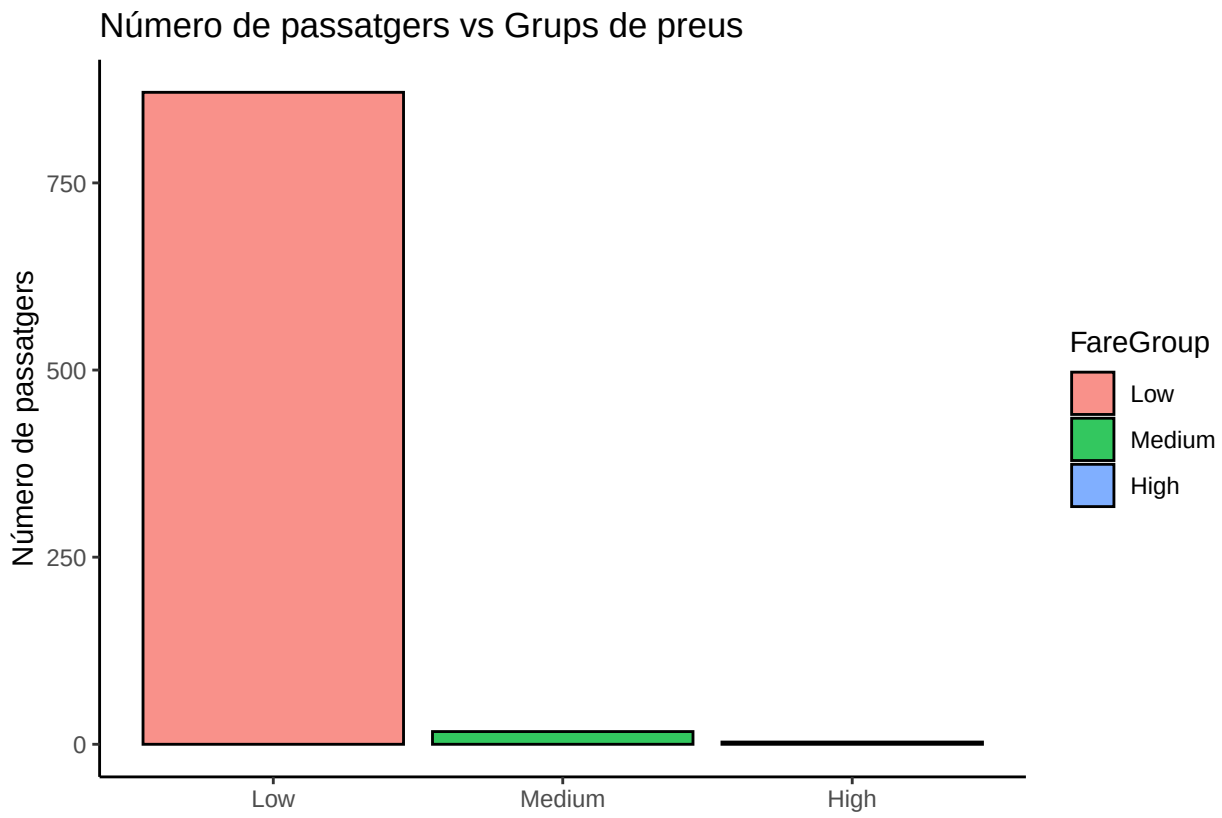


S'aprecia que la majoria dels passatgers viatjaven amb bitllets econòmics. Agrupem ara per grups de preus:

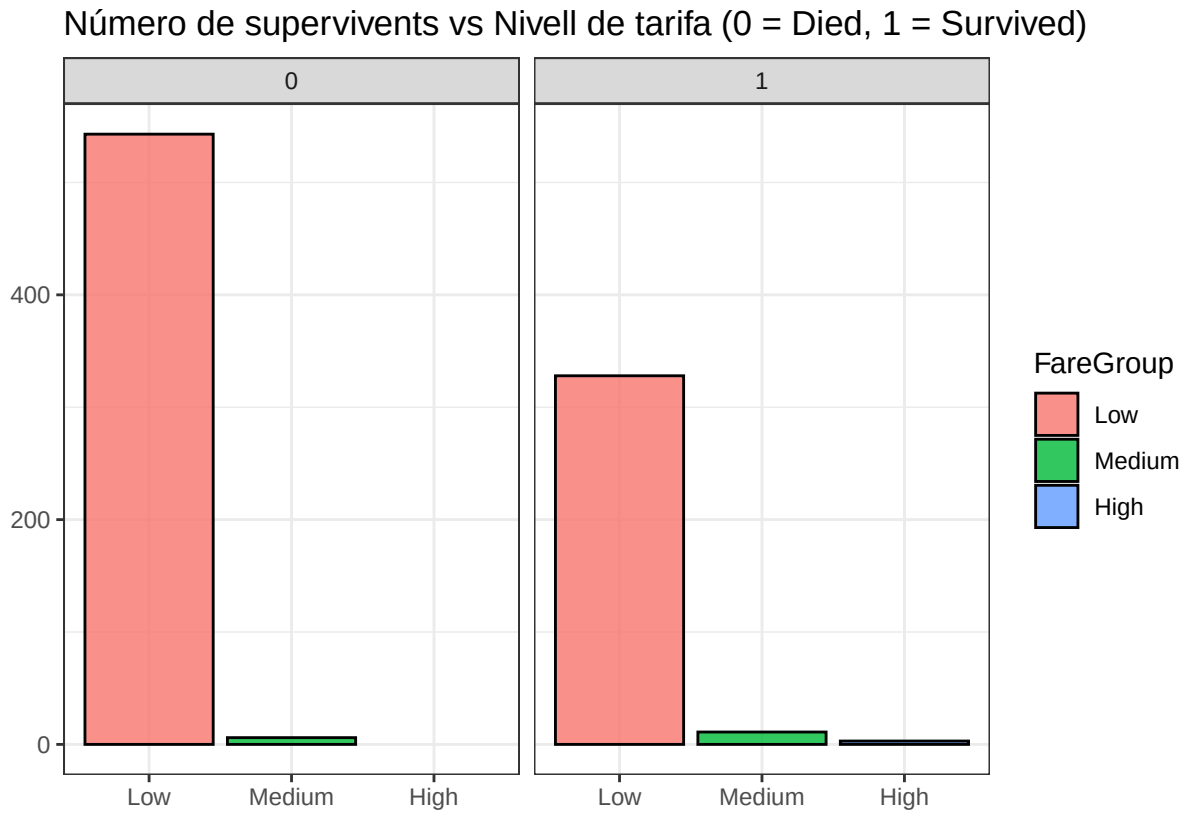
```
faregroupCount <- train %>%
    group_by(FareGroup) %>%
    count(FareGroup) %>%
    select(FareGroup, Passengers = n)

pandoc.table(faregroupCount)
```

```
##
## ------------------------
##  FareGroup   Passengers
## ----------- ------------
##     Low          871
##
##   Medium          17
##
##    High           3
## ------------------------
```

```
ggplot(faregroupCount, aes(x = FareGroup, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = FareGroup), col = "black", alpha = 0.8) +
    labs(x = "", y = "Número de passatgers", title = "Número de passatgers vs Grups de preus") +
    theme_classic()
```




```
faregroupSurvived <- train %>%
    group_by(FareGroup) %>%
    count(Survived) %>%
    select(FareGroup, Survived, Passengers = n)

ggplot(faregroupSurvived, aes(x = FareGroup, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = FareGroup), col = "black", alpha = 0.8) +
    labs(x = "", y = "", title = "Número de supervivents vs Nivell de tarifa (0 = Died, 1 = Survived)") +
    facet_wrap(~Survived) +
    theme_bw()
```



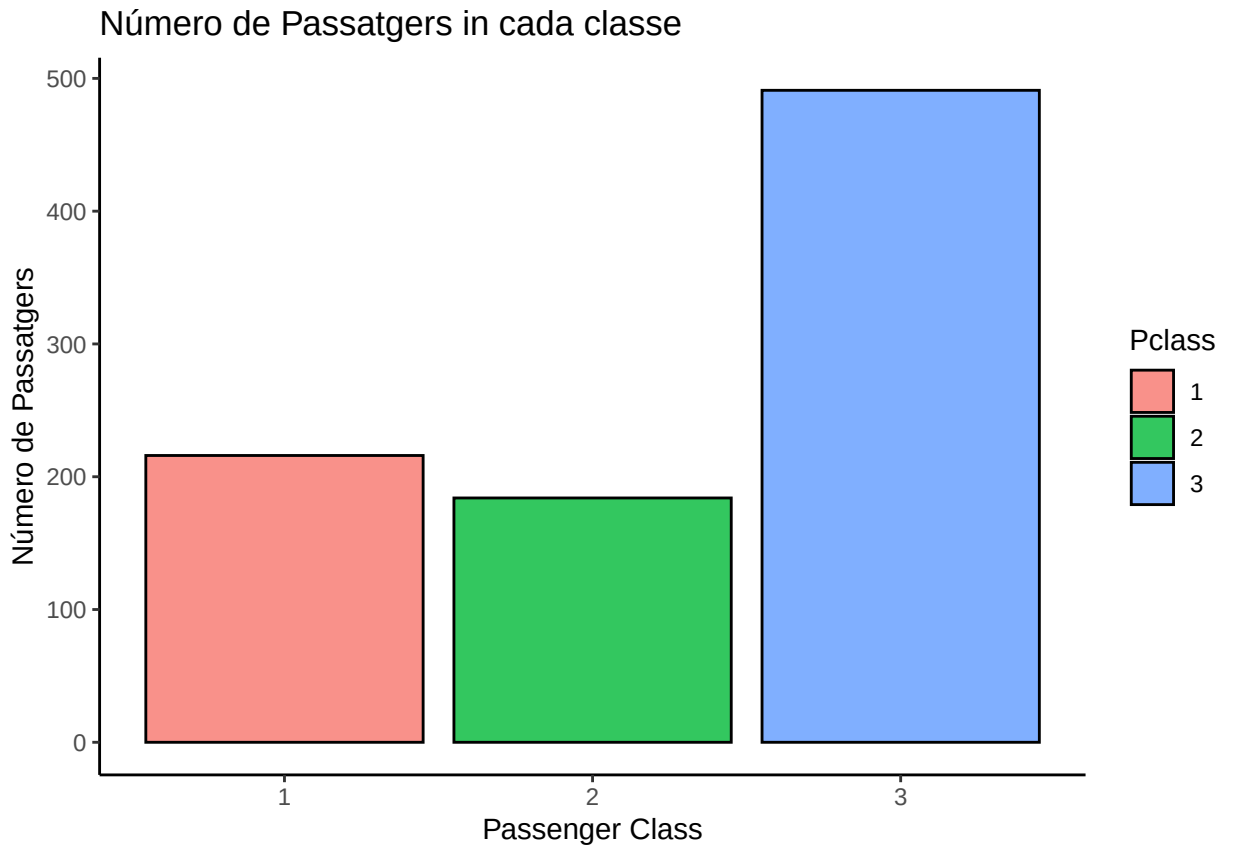
Es pot visualitzar en les gràfiques com les possibilitats de sobreviure augmentaven amb el preu del bitllet del passatger.

Analitzem ara l'impacte de la classe sobre la supervivència:

```
classCount <- train %>%
    group_by(Pclass) %>%
    count(Pclass) %>%
    select(Pclass, Passengers = n)

pandoc.table(classCount)
```

```
##
## ---------------------
##  Pclass   Passengers
## -------- ------------
##    1         216
##
##    2         184
##
##    3         491
## ---------------------
```

```
ggplot(classCount, aes(x = Pclass, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = Pclass), col = "black", alpha = 0.8) +
    labs(x = "Passenger Class", y = "Número de Passatgers", title = "Número de Passatgers in cada classe
    theme_classic()
```




```
classSurvived <- train %>%
    group_by(Pclass) %>%
    count(Survived) %>%
    filter(Survived == 1) %>%
    select(Pclass, Passengers = n)

pandoc.table(classSurvived)
```

```
##
## ---------------------
##  Pclass   Passengers
## -------- ------------
##    1          136
##
##    2           87
##
##    3          119
## ---------------------
```

```
classSurvived <- train %>%
    group_by(Pclass) %>%
    count(Survived) %>%
    select(Pclass, Survived, Passengers = n)

ggplot(classSurvived, aes(x = Pclass, y = Passengers)) +
```

```
    geom_bar(stat = "identity", aes(fill = Pclass), col = "black", alpha = 0.8) +
    labs(x = "", y = "Número de Passatgers", title = "Número de supervivents vs Classe del passatger (0
    facet_wrap(~Survived) +
    theme_bw()
```

Com en l'anàlisi anterior, es veu que els passatgers que viatjaven en primera classe tenien una expectativa de sobreviure (aprox. 65%) superior als de segona classe (aprox. 50%) i molt més elevada que els de tercera classe (aprox. 25%).

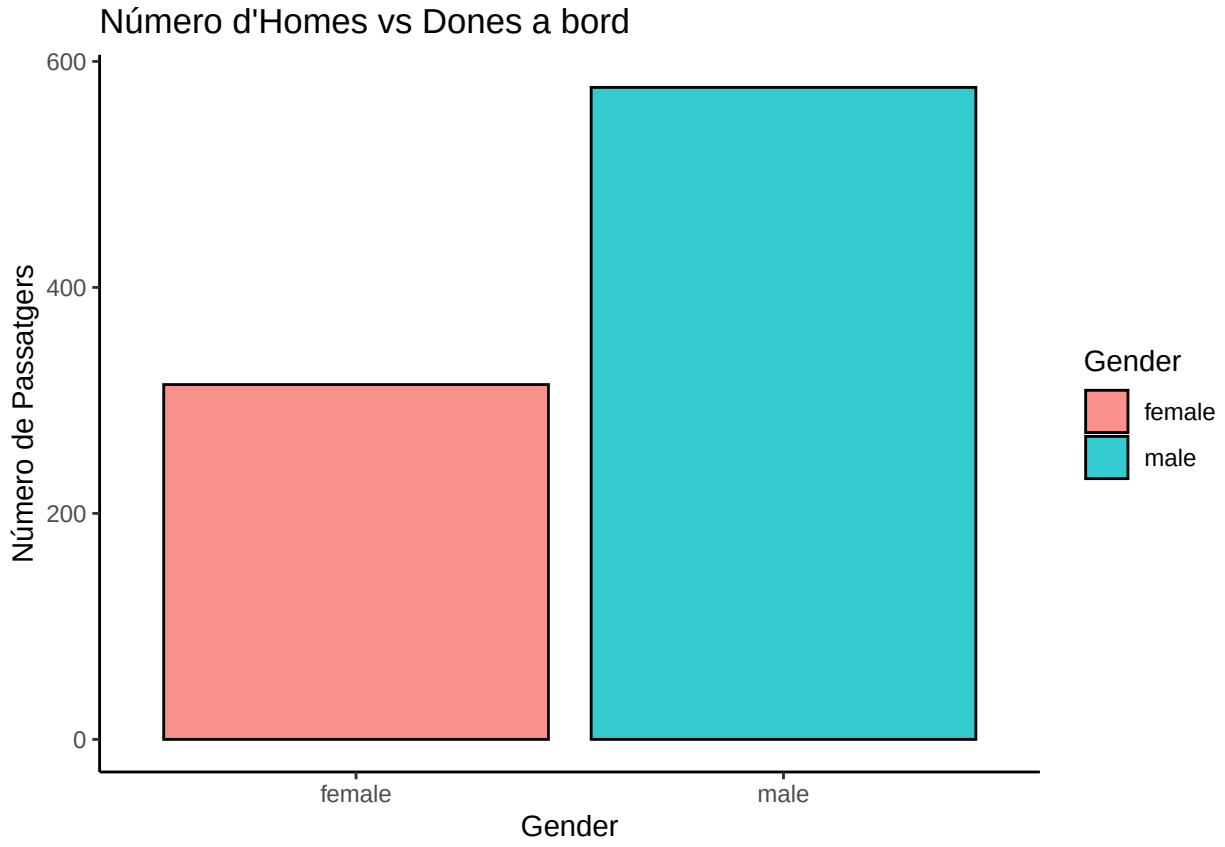
Pel que fa a l'impacte del gènere del passatger en la seva expectativa de supervivència, repetim el mateix anàlisi usant el camp Sex:

```
sexCount <- train %>%
    group_by(Sex) %>%
    count(Sex) %>%
    select(Gender = Sex, Passengers = n)

pandoc.table(sexCount)
```

```
##
## --------------------
##  Gender   Passengers
## -------- ------------
##  female      314
##
##   male       577
## --------------------
```

D'entrada veiem que hi havia força més homes que dones.

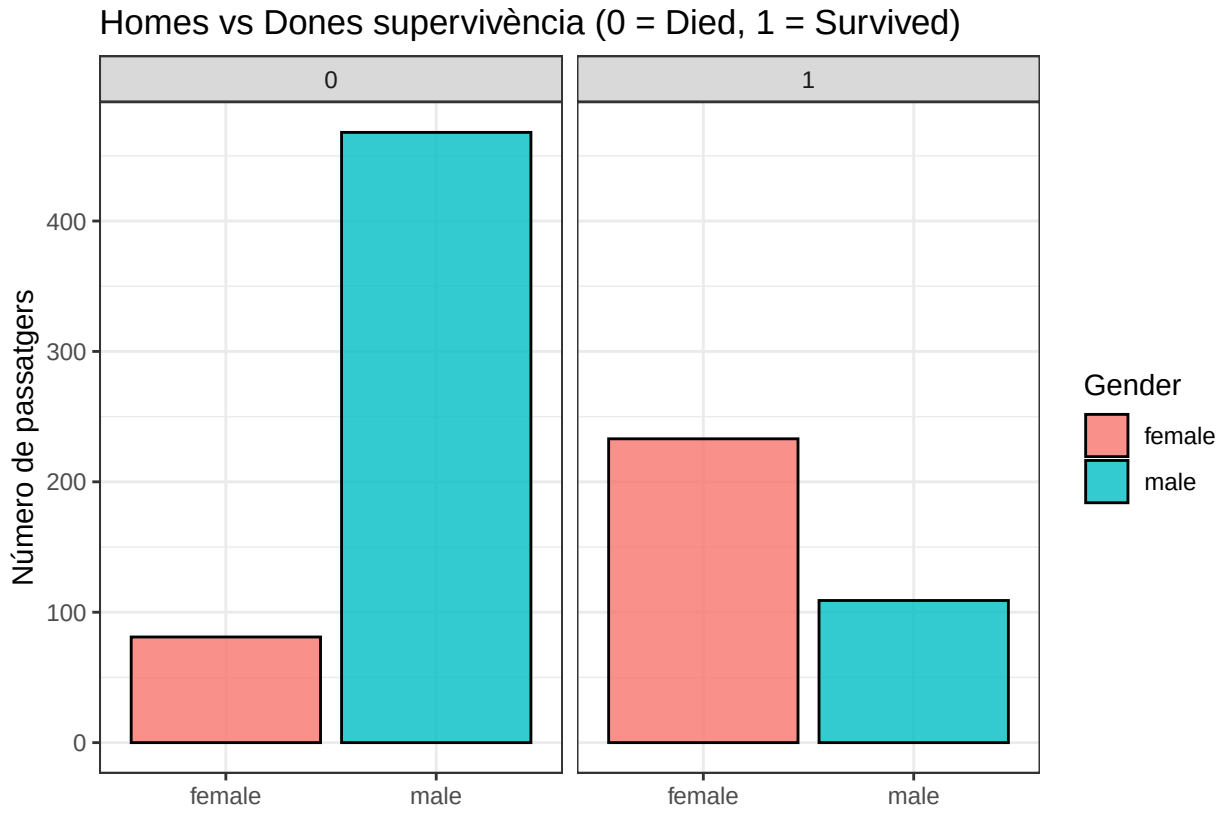
```
ggplot(sexCount, aes(x = Gender, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = Gender), col = "black", alpha = 0.8) +
    labs(x = "Gender", y = "Número de Passatgers", title = "Número d'Homes vs Dones a bord") +
    theme_classic()
```




```
sexSurvived <- train %>%
    group_by(Sex) %>%
    count(Survived) %>%
    filter(Survived == 1) %>%
    select(Gender = Sex, Passengers = n)

pandoc.table(sexSurvived)
```

```
##
## --------------------
##  Gender   Passengers
## -------- ------------
##  female       233
##
##   male        109
## --------------------
```

```
sexSurvived <- train %>%
    group_by(Sex) %>%
    count(Survived) %>%
    select(Gender = Sex, Survived, Passengers = n)

ggplot(sexSurvived, aes(x = Gender, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = Gender), col = "black", alpha = 0.8) +
    labs(x = "", y = "Número de passatgers", title = "Homes vs Dones supervivència (0 = Died, 1 = Survi
    facet_wrap(~Survived) +
    theme_bw()
```



En aquest cas es constata que la gran majoria de dones (al voltant del 75%) que viatjaven en el Titanic van sobreviure. En canvi, la gran majoria d'homes (aprx. 80%) passatgers NO van sobreviure.

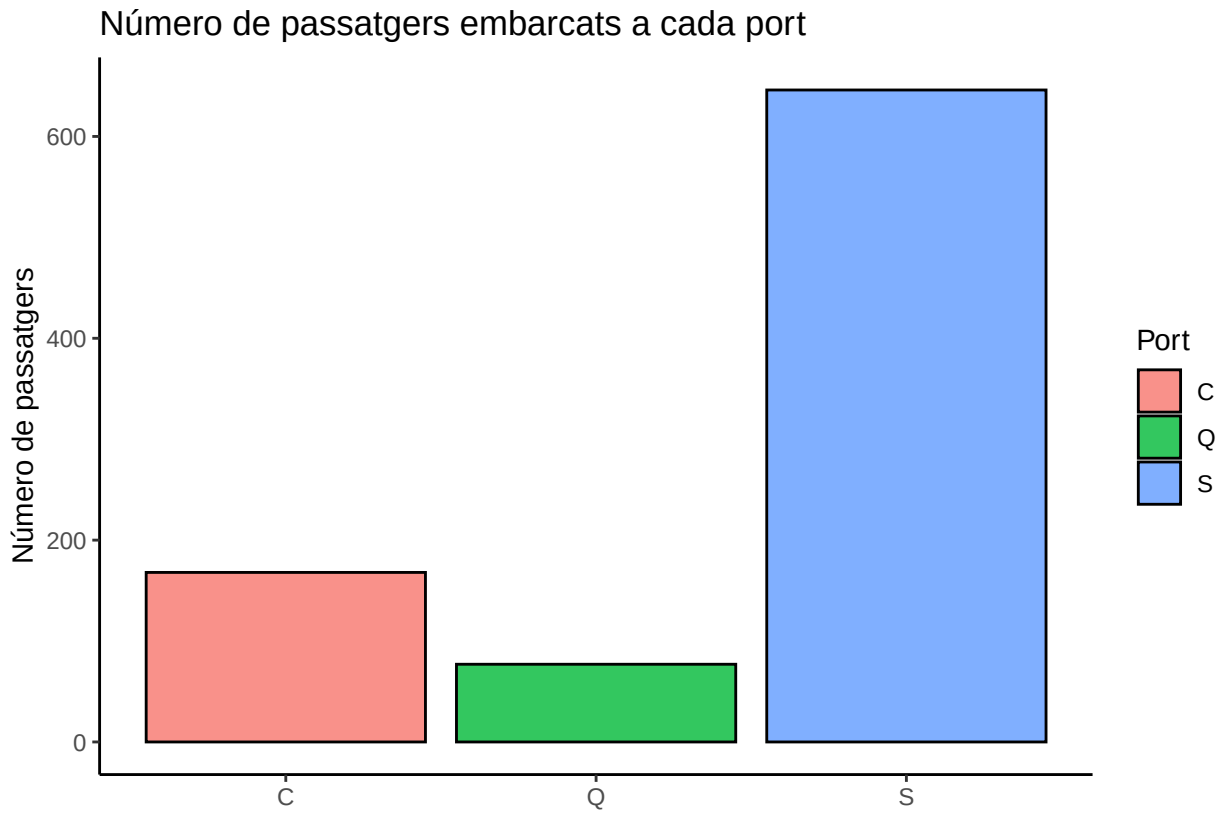
Analitzem a continuació la relació entre el port d'embarcament i la supervivència dels passatgers:

```
embarkedCount <- train %>%
    group_by(Embarked) %>%
    count(Embarked) %>%
    select(Port = Embarked, Passengers = n)

pandoc.table(embarkedCount)
```

```
##
## ------------------
##  Port   Passengers
```

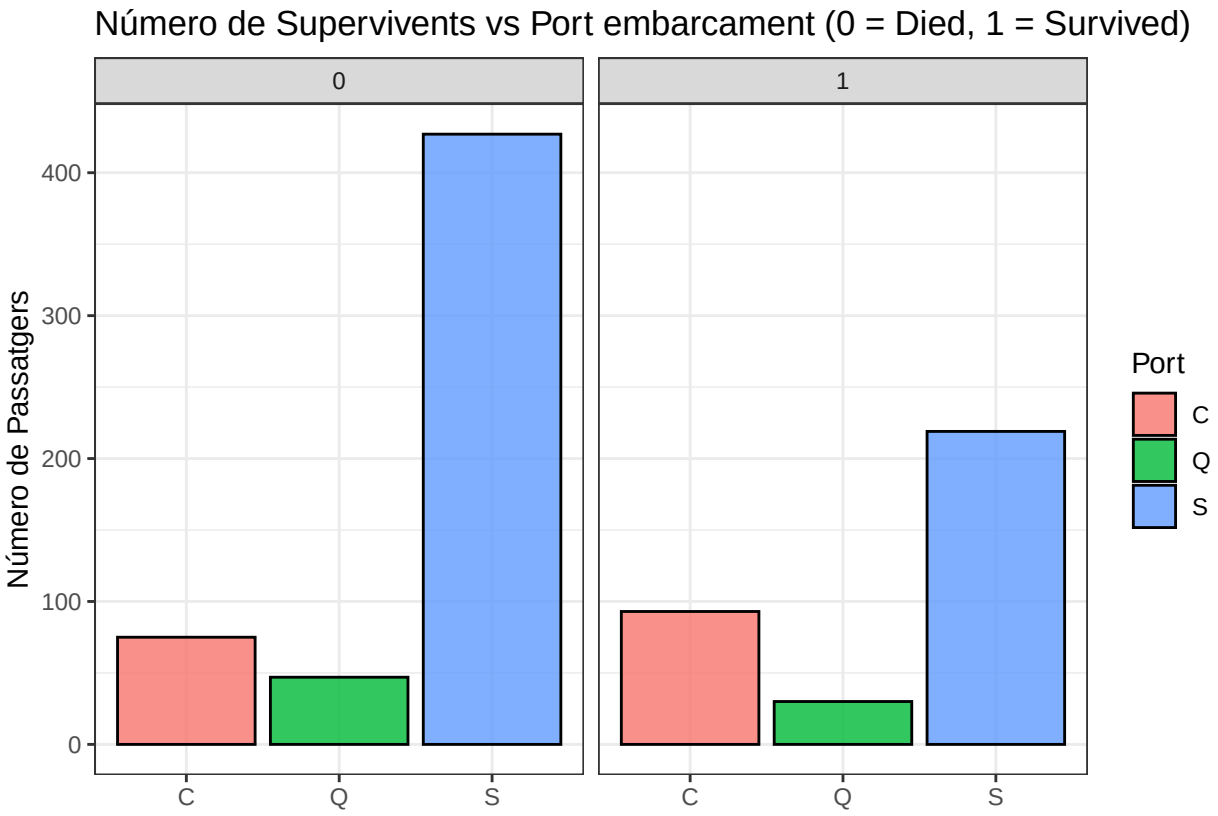
```
## ------ ------------
##   C         168
##
##   Q          77
##
##   S         646
## -------------------
```

```
ggplot(embarkedCount, aes(x = Port, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = Port), col = "black", alpha = 0.8) +
    labs(x = "", y = "Número de passatgers", title = "Número de passatgers embarcats a cada port") +
    theme_classic()
```



La gran majoria dels passatgers provenen del port de Southampton.

```
embarkedSurvived <- train %>%
    group_by(Embarked) %>%
    count(Survived) %>%
    select(Port = Embarked, Survived, Passengers = n)

ggplot(embarkedSurvived, aes(x = Port, y = Passengers)) +
    geom_bar(stat = "identity", aes(fill = Port), col = "black", alpha = 0.8) +
    labs(x = "", y = "Número de Passatgers", title = "Número de Supervivents vs Port embarcament (0 = D
    facet_wrap(~Survived) +
    theme_bw()
```



Els passatgers embarcats a Cherbourg (C) tenen una probabilitat de supervivència superior (aprox. 55%) als provinents dels altres dos ports (Q, aprox. 40%; S, aprox.35%)

Ara podem provar la hipòtesi de que els supervivents eren més joves que els que van morir.

Hipòtesi nul · la i alternativa

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 > p_2$$

Siguent p_1 la proporció de dones supervivents i p_2 la proporció d'homes supervivents.

```
x1 <- train$Sex[train$Sex == 'female' & train$Survived == 1]
x2 <- train$Sex[train$Sex == 'male' & train$Survived == 1]

n1 <- length(x1)
n2 <- length(x2)

p1 <- n1 / (n1 + n2)
p2 <- n2 / (n1 + n2)

p <- (n1*p1 + n2*p2) / (n1+n2)

success <- c( p1*n1, p2*n2)
```

```
nn <- c(n1,n2)

prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  success out of nn
## X-squared = 39.735, df = 1, p-value = 1.454e-10
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2736296 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.6812865 0.3187135
```

Com el p-value és inferior a 0.05 podem rebutjar la hipòtesi nul · la i concloure com ja esperavem que la proporció de dones supervivents és major que la d'homes.

Predicció

Després d'explorar les dades, anem a intentar construïr un model de predicció.

```
library(caret)
library(randomForest)
```

Dividim les dades de train en 70% training i 30% test:

```
set.seed(1995)
inTrain <- createDataPartition(train$Survived, p = 0.7, list = FALSE)
training <- train[inTrain, ]
testing <- train[-inTrain, ]
```

Construïm el model utilitzant un Random Forest:

```
set.seed(1995)
rfModel <-
  randomForest(Survived ~ Pclass + Sex + AgeGroup + FareGroup + Embarked, data = training)
rfPred <- predict(rfModel, newdata = testing)
rfCM <- confusionMatrix(rfPred, testing$Survived)
rfCM$table
```

```
##            Reference
## Prediction   0   1
##          0 157  46
##          1   7  56
```

Mirem el percentatge d'acert en la predicció:

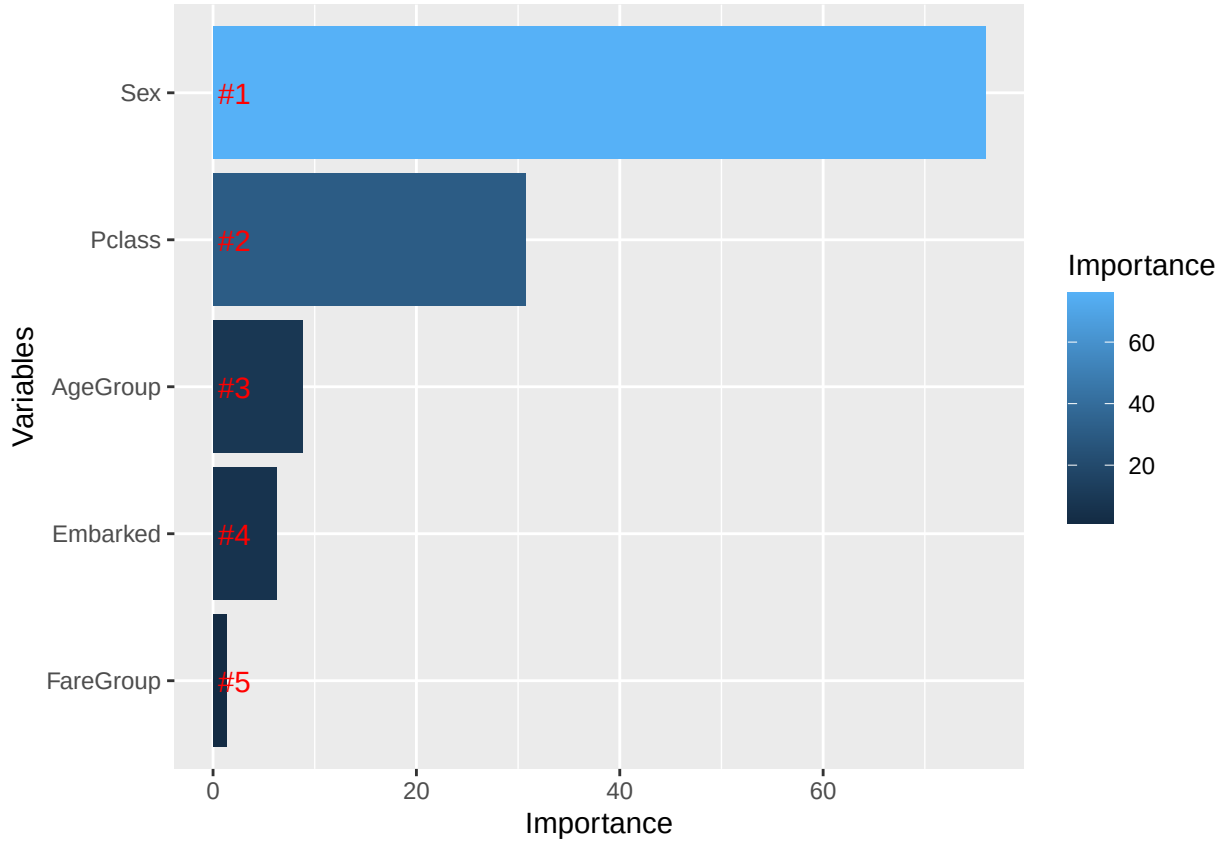
```r
accuracy <- rfCM$overall[1]
accuracy
```

```
##  Accuracy
## 0.8007519
```

Obtenim un 80% d'acert en la predicció.

Si mirem la importància de cadascuna de les variables:

```r
# Obtenim la importància
importance    <- importance(rfModel)
varImportance <- data.frame(Variables = row.names(importance),
                            Importance = round(importance[ ,'MeanDecreaseGini'],2))

# Creem una variable de rang basada en la importància
rankImportance <- varImportance %>%
  mutate(Rank = paste0('#',dense_rank(desc(Importance))))

# Utilitzem ggplo2 per a visualitzar la importància relativa de les variables
ggplot(rankImportance, aes(x = reorder(Variables, Importance),
    y = Importance, fill = Importance)) +
  geom_bar(stat='identity') +
  geom_text(aes(x = Variables, y = 0.5, label = Rank),
    hjust=0, vjust=0.55, size = 4, colour = 'red') +
  labs(x = 'Variables') +
  coord_flip()
```



Com vegem el sexe de la persona té més influència que la classe!

Ara podem utilitzar aquest model per a fer la predicció sobre les dades originals de test i guardar el resultat a un fitxer csv:

```
prediction <- predict(rfModel, newdata = test)

titanicPrediction <-
  data.frame(PassengerId = test$PassengerId, Survived = prediction)
write.csv(titanicPrediction, file = "TitanicPrediction.csv", row.names = FALSE)
```