PAC2

Felix Ribera Manel Benavides

5/1/2021

Lectura de dades

Es carreguen les dades i es comprova que els tipus de les dades siguin els esperats.

```
df <- read.csv("train.csv")
head(df)
```

```
##   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
```

```
str(df)
```

```
## '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)"
##  $ 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" ...
```

```
summary(df)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Hi ha tres variables que no ens aporten informació a nivell estadístic a l'hora d'analitzar les dades. Aquestes variables són: **Name**, **Ticket** i **PassengerId**. Per tant, aquestes variables seran eliminades de cara a l'anàlisi.

```
df <- select(df, -Name)
df <- select(df, -Ticket)
df <- select(df, -PassengerId)
```

Es passen els valors categòrics a tipus factor.

```
df$Pclass <- as.factor(df$Pclass)
df$Sex <- as.factor(df$Sex)
df$Embarked <- as.factor(df$Embarked)

str(df)
```

```
## 'data.frame':    891 obs. of  9 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
```

```
##  $ 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 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin   : chr  "" "C85" "" "C123" ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

Es comprova si hi ha valors NA en el dataset, o valors buits.

```
colSums(is.na(df))
```

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

```
colSums(df=="")
```

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

S'observa que hi ha 177 valors d'Age NA, 687 Cabin buits i 2 Embarked.

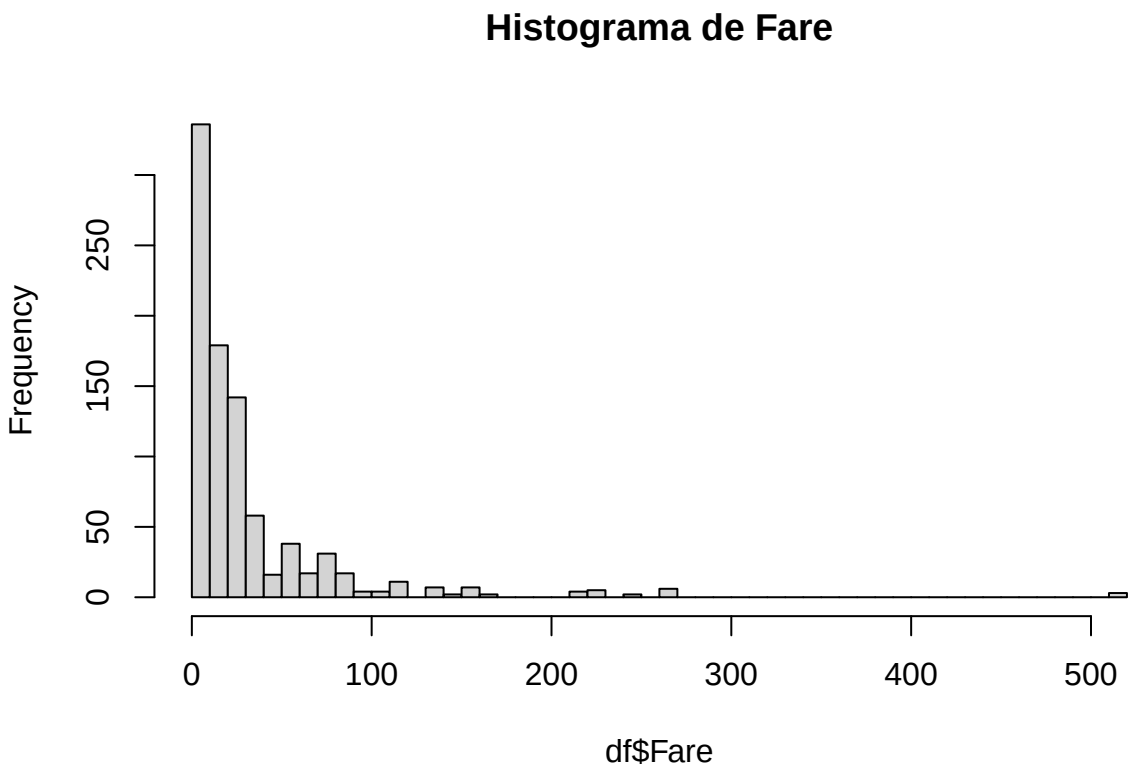
Posem el valor de la mitjana en els missing values d'Age.

```
age_mean <- mean(df$Age[!is.na(df$Age)])
df$Age[is.na(df$Age)] <- age_mean
```

Identificació i tractament de valors extrems.

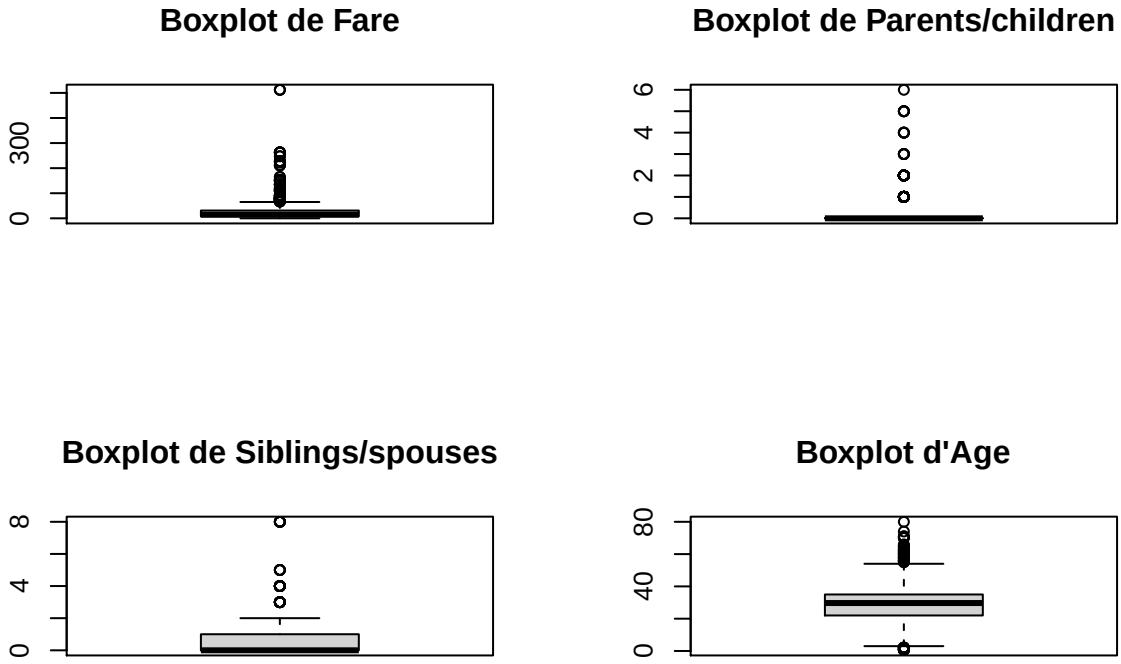
Analitzem la distribució dels valors de Fare.

```
hist(df$Fare, breaks=50, main="Histograma de Fare")
```



La majoria de valors es troben en el grup < 150 i majoritàriament es troben en el grup < 50.

```
par(mfrow=c(2,2))
boxplot(df$Fare, main="Boxplot de Fare")
boxplot(df$Parch, main="Boxplot de Parents/children")
boxplot(df$SibSp, main="Boxplot de Siblings/spouses")
boxplot(df$Age, main="Boxplot d'Age")
```



Com hem vist en l'histograma anteriorment hi ha outliers en els valors de Fare. Tot i això considerem que no son valors erronis, sino simplement elevats. Es pot deure a habitacions del Titanic més exclusives.

Per la resta de variables veiem que existeixen outliers, però són valors raonables que es deuen a un major número mostres de certs valors i no a errors.

Neteja de la variable Cabin

A cabin hi veiem representat un caràcter amb un número. S'ha considerat interessant separar aquesta informació per veure com afecta tant la ubicació dins el vaixell (representada pel caràcter) com el número d'habitació.

```
f_split_cabin <- function (x) {
  cabin <- strsplit(x, " ")[[1]][1]
  type <- substring(cabin, 1, 1)
  number <- substring(cabin, 2)
  if (is.na(type)) type <- ""
  if (is.na(number) || number == "") number <- 0
  return(c(type, number))
}

cabinType <- c()
cabinNumber <- c()
for (item in df$Cabin){
  cabin <- f_split_cabin(item)
  cabinType <- c(cabinType, cabin[1])
```

```
  cabinNumber <- c(cabinNumber, cabin[2])
}

df["cabinType"] <- as.factor(cabinType)
df["cabinNumber"] <- as.integer(cabinNumber)

df <- select(df, -Cabin)

head(df)
```

```
##   Survived Pclass    Sex      Age SibSp Parch    Fare Embarked cabinType
## 1        0      3   male 22.00000     1     0  7.2500        S
## 2        1      1 female 38.00000     1     0 71.2833        C         C
## 3        1      3 female 26.00000     0     0  7.9250        S
## 4        1      1 female 35.00000     1     0 53.1000        S         C
## 5        0      3   male 35.00000     0     0  8.0500        S
## 6        0      3   male 29.69912     0     0  8.4583        Q
##   cabinNumber
## 1           0
## 2          85
## 3           0
## 4         123
## 5           0
## 6           0
```

Anàlisi de les dades

```
# Agrupació per classe.
df.p1 <- df[df$Pclass == "1",]
df.p2 <- df[df$Pclass == "2",]
df.p3 <- df[df$Pclass == "3",]

#Agrupació per port d'embarc
df.emb_s <- df[df$Embarked == "S",]
df.emb_q <- df[df$Embarked == "Q",]
df.emb_c <- df[df$Embarked == "C",]

# Agrupació per cabinType
df.ctype_a <- df[df$cabinType == "A",]
df.ctype_b <- df[df$cabinType == "B",]
df.ctype_c <- df[df$cabinType == "C",]
df.ctype_d <- df[df$cabinType == "D",]
df.ctype_e <- df[df$cabinType == "E",]
df.ctype_f <- df[df$cabinType == "F",]
df.ctype_g <- df[df$cabinType == "G",]
df.ctype_t <- df[df$cabinType == "T",]

# Agrupació per sex
df.male <- df[df$Sex == "male",]
df.female <- df[df$Sex == "female",]
```

Comprovació de la normalitat i homogenitat

```
alpha = 0.05
col.names = colnames(df)

for (i in 1:ncol(df)) {
  if (is.integer(df[,i]) | is.numeric(df[,i])) {
    p_val = ad.test(df[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      cat('\n')
    }
  }
}
```

```
## Survived
## Age
## SibSp
## Parch
## Fare
## cabinNumber
```

```
fligner.test(Survived ~ Age, data = df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Age
## Fligner-Killeen:med chi-squared = 76.02, df = 88, p-value = 0.8151
```

```
fligner.test(Survived ~ Fare, data = df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Fare
## Fligner-Killeen:med chi-squared = 258.22, df = 247, p-value = 0.299
```

```
fligner.test(Survived ~ SibSp, data = df)
```

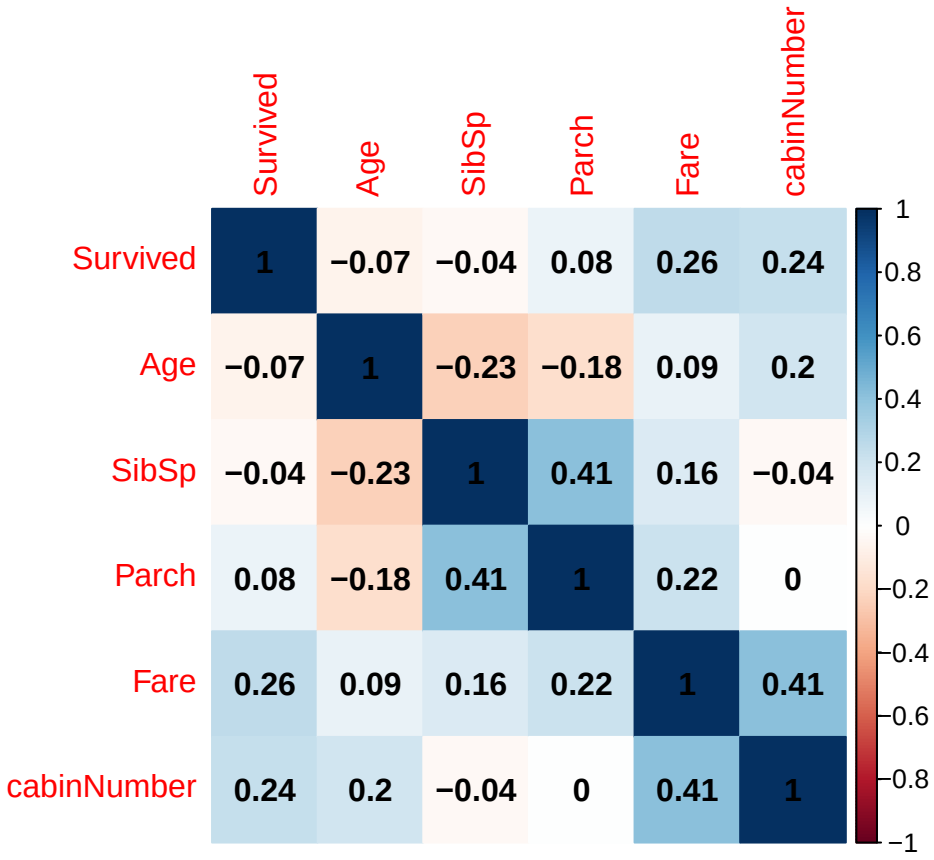
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by SibSp
## Fligner-Killeen:med chi-squared = 21.832, df = 6, p-value = 0.001298
```

```
fligner.test(Survived ~ Parch, data = df)
```

```
##
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Parch
## Fligner-Killeen:med chi-squared = 17.231, df = 6, p-value = 0.00847
```

```
df_num <- df[, sapply(df, is.numeric)]

corrplot(cor(df_num), method="color", addCoef.col = "black")
```




```
str(df)
```

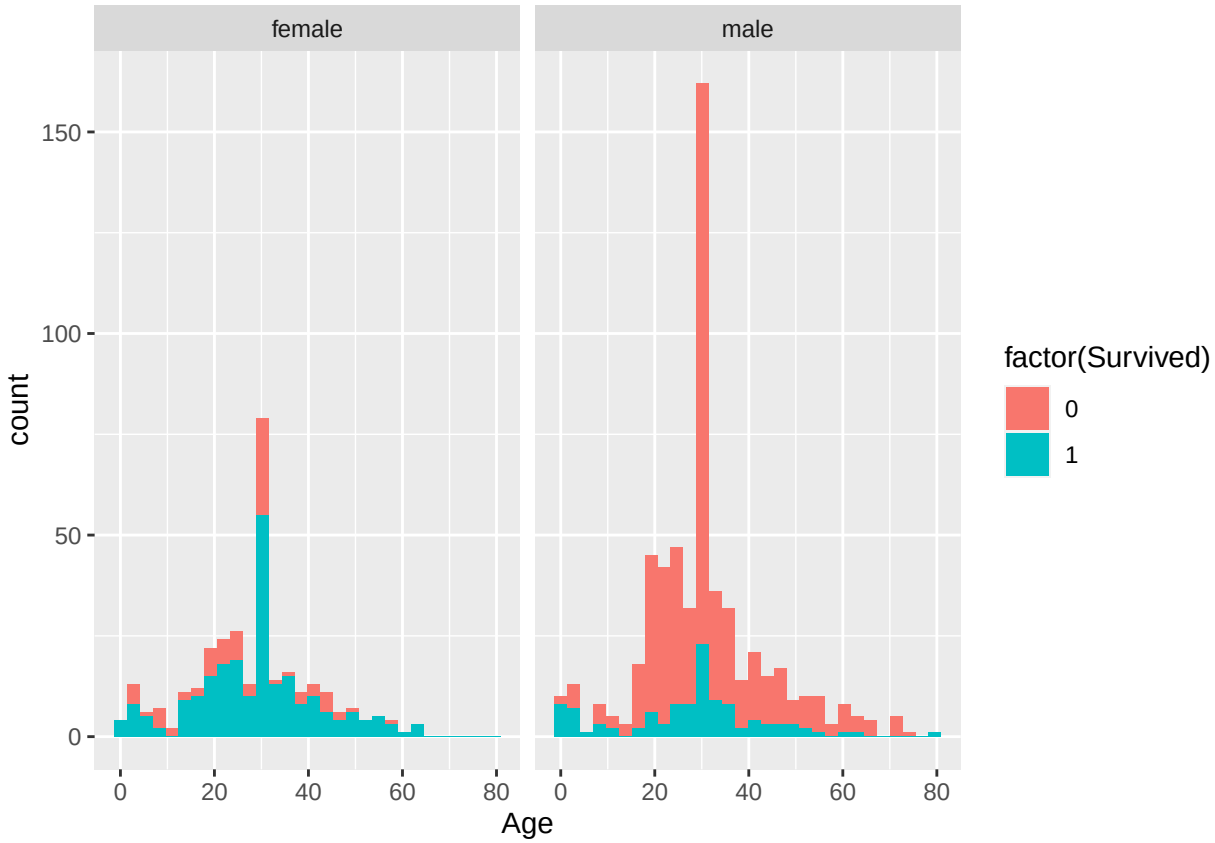
```
## 'data.frame':    891 obs. of  10 variables:
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
##  $ cabinType  : Factor w/ 9 levels "","A","B","C",..: 1 4 1 4 1 1 6 1 1 1 ...
##  $ cabinNumber: int  0 85 0 123 0 0 46 0 0 0 ...
```

```r
df.male.survived <- df[df$Sex == "male",]$Survived
df.female.survived <- df[df$Sex == "female",]$Survived

t.test(df.male.survived, df.female.survived, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  df.male.survived and df.female.survived
## t = -18.672, df = 584.43, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5043259
## sample estimates:
## mean of x mean of y
## 0.1889081 0.7420382
```

```r
ggplot(df[1:891,], aes(Age, fill = factor(Survived))) +
  geom_histogram() +
  facet_grid(.~Sex)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

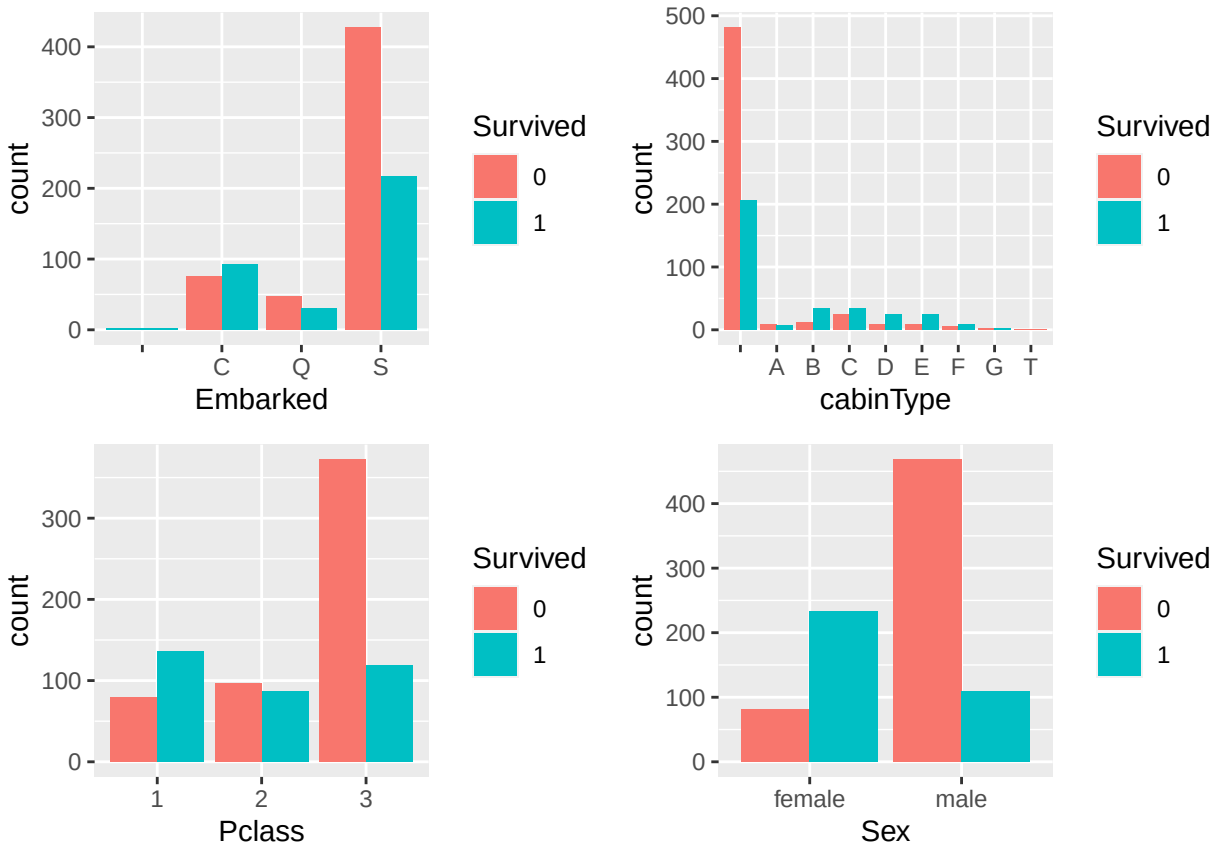


```
model_1 <- lm(Survived ~ Age + Parch + SibSp + Fare + cabinNumber, data = df)
model_2 <- lm(Survived ~ Pclass + Sex + Embarked + cabinType, data = df)
model_3 <- lm(Survived ~ Age + Parch + SibSp + Fare + cabinNumber +
                Pclass + Sex + Embarked + cabinType, data = df)

taula.coeficients <- matrix(c(
  "Quantitatives", summary(model_1)$r.squared,
  "Qualitatives", summary(model_2)$r.squared,
  "Mix", summary(model_3)$r.squared
  ),
    ncol = 2, byrow = TRUE)
colnames(taula.coeficients) <- c("Model", "R^2")
taula.coeficients
```

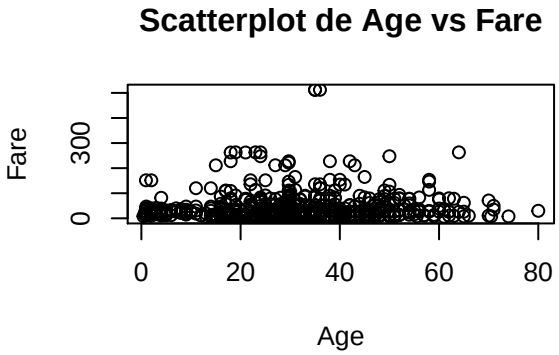
```
##      Model           R^2
## [1,] "Quantitatives" "0.112346236189579"
## [2,] "Qualitatives"  "0.387144392223383"
## [3,] "Mix"           "0.413430912164883"
```

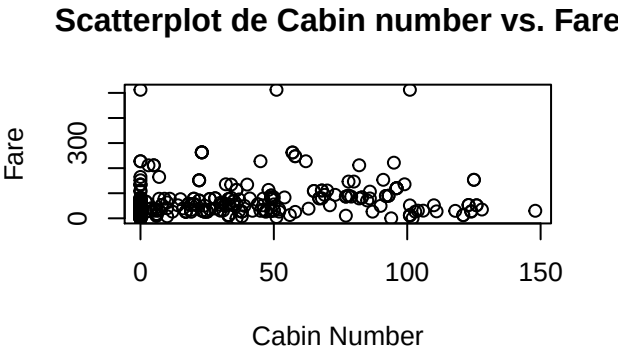
Representacions gràfiques

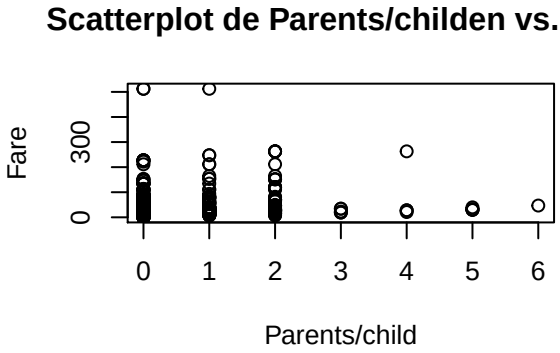
```
gg_em <- ggplot(df, aes(x = Embarked, fill = as.factor(Survived))) +
  labs(fill="Survived") +
  geom_bar(position = "dodge")

gg_ctype <- ggplot(df, aes(x = cabinType, fill = as.factor(Survived))) +
  labs(fill="Survived") +
  geom_bar(position = "dodge")

gg_pclass <- ggplot(df, aes(x = Pclass, fill = as.factor(Survived))) +
  labs(fill="Survived") +
  geom_bar(position = "dodge")

gg_sex <- ggplot(df, aes(x = Sex, fill = as.factor(Survived))) +
  labs(fill="Survived") +
  geom_bar(position = "dodge")

grid.arrange(gg_em, gg_ctype, gg_pclass, gg_sex, nrow=2)
```




```
par(mfrow=c(2,2))

plot(df$Age,df$Fare, main="Scatterplot de Age vs Fare",
     xlab="Age", ylab = "Fare")
plot(df$cabinNumber, df$Fare, main="Scatterplot de Cabin number vs. Fare",
     xlab="Cabin Number", ylab = "Fare")
plot(df$Parc,df$Fare, main="Scatterplot de Parents/childen vs. Fare",
     xlab="Parents/child", ylab = "Fare")
```


Scatterplot de Age vs Fare
Fare
0
300
0
20
40
60
80
Age


Scatterplot de Cabin number vs. Fare
Fare
0
300
0
50
100
150
Cabin Number


Scatterplot de Parents/childen vs. Far
Fare
0
300
0
1
2
3
4
5
6
Parents/child