

# Práctica 2

Marco Emilio Rodríguez Serrano && Luis García Tarraga

5 de enero, 2021

- 0.Carga de los datos
- 1.Descripción del dataset.
- 2.Integración y selección de los datos de interés a analizar.
- 3.Limpieza de los datos.
  - 3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?
  - 3.2. Identificación y tratamiento de valores extremos.
- 4.Análisis de los datos.
  - 4.1.Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).
  - 4.2.Comprobación de la normalidad y homogeneidad de la varianza.
  - 4.3.Aplicación de pruebas estadísticas para comparar los grupos de datos.
    - 4.3.1 Hipótesis edad supervivientes
    - 4.3.2 Hipótesis sexo supervivientes
    - 4.3.2 Regresión con edad
    - 4.3.3 Correlación
    - 4.3.4 Árbol de decisión
- 5.Representación gráfica
- 6.Conclusiones
- 7.Contribuciones

# 0.Carga de los datos

Cargamos los datos con la cabecera

Visualizamos las 5 primeras filas para comprobar de forma visual que no hay problemas aparentes

```
#Cargamos el conjunto de datos
data <- read.csv('train.csv',stringsAsFactors = TRUE, header = TRUE, strip.white = T, sep =
",")  
  
summary(data)
```

```
##   PassengerId      Survived      Pclass
##   Min.   : 1.0   Min.   :0.0000   Min.   :1.000
##   1st Qu.:223.5  1st Qu.:0.0000  1st Qu.:2.000
##   Median :446.0   Median :0.0000  Median :3.000
##   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
##
##                                     Name      Sex      Age
##   Abbing, Mr. Anthony           : 1  female:314  Min.   : 0.42
##   Abbott, Mr. Rossmore Edward   : 1    male :577  1st Qu.:20.12
##   Abbott, Mrs. Stanton (Rosa Hunt) : 1                   Median :28.00
##   Abelson, Mr. Samuel          : 1                   Mean   :29.70
##   Abelson, Mrs. Samuel (Hannah Wizosky): 1               3rd Qu.:38.00
##   Adahl, Mr. Mauritz Nils Martin : 1                   Max.   :80.00
##   (Other)                      :885                   NA's   :177
##   SibSp            Parch      Ticket      Fare
##   Min.   :0.000   Min.   :0.0000  1601   : 7  Min.   : 0.00
##   1st Qu.:0.000   1st Qu.:0.0000  347082 : 7  1st Qu.: 7.91
##   Median :0.000   Median :0.0000  CA. 2343: 7  Median :14.45
##   Mean   :0.523   Mean   :0.3816  3101295: 6  Mean   :32.20
##   3rd Qu.:1.000   3rd Qu.:0.0000  347088 : 6  3rd Qu.:31.00
##   Max.   :8.000   Max.   :6.0000  CA 2144 : 6  Max.   :512.33
##                           (Other) :852
##   Cabin      Embarked
##   :687      : 2
##   B96 B98   : 4  C:168
##   C23 C25 C27: 4  Q: 77
##   G6        : 4  S:644
##   C22 C26   : 3
##   D         : 3
##   (Other)   :186
```

```
head(data, 5)
```

```
## PassengerId Survived Pclass
## 1          1      0      3
## 2          2      1      1
## 3          3      1      3
## 4          4      1      1
## 5          5      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
##             Ticket   Fare Cabin Embarked
## 1         A/5 21171  7.2500          S
## 2            PC 17599 71.2833        C85
## 3  STON/O2. 3101282  7.9250          S
## 4        113803 53.1000        C123
## 5        373450  8.0500          S
```

# 1. Descripción del dataset.

Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset se compone de los siguientes campos:

- **PassengerId**: Identificador del pasajero
- **Survived**: Indica si el pasajero sobrevivió. Si vale 0 entonces no sobrevivió, si vale 1 entonces es un superviviente
- **Pclass**: Indica la clase en la que viajaba el pasajero (1 = 1st, 2 = 2nd, 3 = 3rd)
- **Name**: Nombre y apellidos del pasajero
- **Sex**: Sexo del pasajero
- **Age**: Edad en años
- **SibSp**: Número de hermanos / esposas a bordo del Titanic
- **Parch**: Número de padres / hijos a bordo del Titanic
- **Ticket**: Número de ticket
- **Fare**: Precio que ha pagado el pasajero por el viaje
- **Cabin**: Número de la cabina del pasajero
- **Embarked**: Puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton)

Este dataset es muy relevante porque se refiere a los pasajeros del Titanic, indicando además si el pasajero consiguió sobrevivir o no al naufragio.

A través de este dataset se puede analizar distintos aspectos sobre quiénes sobrevivieron, por ejemplo si las mujeres y los niños se salvaron en relación a los hombres, si el hecho de viajar en una clase u otra podría implicar un mayor o menor ratio de supervivencia.

Este dataset es además un clásico a nivel de formación y ejemplos de estadística y minería de datos. Este ejemplo en concreto lo hemos obtenido de: <https://www.kaggle.com/c/titanic> (<https://www.kaggle.com/c/titanic>)

## 2. Integración y selección de los datos de interés a analizar.

Desde la vista de resumen de nuestros datos podemos ver que tenemos diversos puntos a revisar, primero tenemos la variable **Name** que tiene una varianza tal que cada registro contiene un valor único, por lo cual esta variable no tiene uso práctico tal cual, sin transformar, por lo que la vamos a eliminar del conjunto de datos. Ocurre lo mismo con **PassengerId**, no necesitamos este campo ya que contiene un número de pasajero pero no aporta información para el análisis.

```
data$Name <- NULL  
  
data$PassengerId <- NULL  
  
summary(data)
```

```
##      Survived          Pclass           Sex            Age          SibSp  
##  Min.   :0.0000  Min.   :1.000  female:314  Min.   : 0.42  Min.   :0.000  
##  1st Qu.:0.0000  1st Qu.:2.000  male   :577  1st Qu.:20.12  1st Qu.:0.000  
##  Median :0.0000  Median :3.000                    Median :28.00  Median :0.000  
##  Mean    :0.3838  Mean   :2.309                    Mean   :29.70  Mean   :0.523  
##  3rd Qu.:1.0000  3rd Qu.:3.000                    3rd Qu.:38.00  3rd Qu.:1.000  
##  Max.   :1.0000  Max.   :3.000                    Max.   :80.00  Max.   :8.000  
##                                         NA's   :177  
##      Parch          Ticket          Fare          Cabin          Embarked  
##  Min.   :0.0000  1601   : 7  Min.   :  0.00       :687   : 2  
##  1st Qu.:0.0000  347082 : 7  1st Qu.:  7.91  B96 B98   : 4  C:168  
##  Median :0.0000  CA. 2343: 7  Median : 14.45  C23 C25 C27: 4  Q: 77  
##  Mean   :0.3816  3101295: 6  Mean   : 32.20  G6        : 4  S:644  
##  3rd Qu.:0.0000  347088 : 6  3rd Qu.: 31.00  C22 C26   : 3  
##  Max.   :6.0000  CA 2144 : 6  Max.   :512.33  D        : 3  
##                  (Other) :852  (Other)       :186
```

### 3. Limpieza de los datos.

Podemos ver que la variable **Survived** es una variable categórica, pero ha cargado como variable numérica, lo mismo sucede con la variable **Pclass**, vamos a transformarlas:

```
data$Survived <- factor(data$Survived, levels = c(0,1), labels = c("NO", "YES"))

data$Pclass <- factor(data$Pclass, levels = c(1,2,3), labels = c("1st", "2nd", "3rd"))

# Revisamos La estructura
str(data)
```

```
## 'data.frame': 891 obs. of 10 variables:
## $ Survived: Factor w/ 2 levels "NO","YES": 1 2 2 2 1 1 1 1 2 2 ...
## $ Pclass   : Factor w/ 3 levels "1st","2nd","3rd": 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 ...
## $ Ticket   : Factor w/ 681 levels "110152","110413",...: 524 597 670 50 473 276 86 396 345
## $ Fare     : num  7.25 71.28 7.92 53.1 8.05 ...
## $ Cabin    : Factor w/ 148 levels "", "A10", "A14", ...: 1 83 1 57 1 1 131 1 1 1 ...
## $ Embarked: Factor w/ 4 levels "", "C", "Q", "S": 4 2 4 4 4 3 4 4 4 2 ...
```

A continuación, vamos a recodificar como NA los valores faltantes en la variable **Cabin** y **Embarked**. Estos datos los trataremos más adelante.

```
data$Cabin[data$Cabin == ""] <- NA
data$Embarked[data$Embarked == ""] <- NA

summary(data)
```

```
##   Survived Pclass      Sex       Age      SibSp
## NO :549   1st:216   female:314   Min.   :0.42   Min.   :0.000
## YES:342   2nd:184   male  :577   1st Qu.:20.12  1st Qu.:0.000
##          3rd:491                    Median :28.00  Median :0.000
##                               Mean   :29.70  Mean   :0.523
##                               3rd Qu.:38.00 3rd Qu.:1.000
##                               Max.   :80.00  Max.   :8.000
##                               NA's    :177
##   Parch      Ticket     Fare      Cabin     Embarked
## Min.   :0.0000  1601   : 7  Min.   : 0.00  B96  B98   : 4   :
## 1st Qu.:0.0000  347082 : 7  1st Qu.: 7.91  C23  C25   : 4   C   :168
## Median :0.0000  CA. 2343: 7  Median :14.45  G6    : 4   Q   : 77
## Mean   :0.3816  3101295: 6  Mean   :32.20  C22  C26   : 3   S   :644
## 3rd Qu.:0.0000  347088 : 6  3rd Qu.:31.00  D    : 3   NA's: 2
## Max.   :6.0000  CA 2144 : 6  Max.   :512.33 (Other) :186
##                   (Other) :852                      NA's   :687
```

A continuación, vamos a recodificar los datos de la cabina para generar una nueva variable conteniendo las plantas en las que residían los pasajeros, que es un dato que, por su menor varianza y mayor relación, aparente, con la variable respuesta puede ser más interesante:

```
n <- nrow(data)
aux <- 0
planta <- c()
for(i in 1:n){
  if(is.na(data$Cabin[i])){planta[i] <- NA}
  else{
    aux <- substr(data$Cabin[i],start = 2, stop = 2)
    if(aux == " ") {planta[i] <- substr(data$Cabin[i], start = 1, stop = 3)}
    else{planta[i] <- substr(data$Cabin[i], start = 1, stop = 1)}
  }
}

data$planta <- planta <- as.factor(factor(planta, levels = c("A", "B", "C", "D", "E", "F", "F E", "F G", "G", "T"), labels = c("A", "B", "C", "D", "E", "F", "F E", "F G", "G", "T")))

summary(data)
```

```
##   Survived Pclass      Sex       Age     SibSp
## NO :549   1st:216 female:314   Min.   :0.42  Min.   :0.000
## YES:342   2nd:184 male  :577   1st Qu.:20.12 1st Qu.:0.000
##                   3rd:491   Median :28.00  Median :0.000
##                               Mean   :29.70  Mean   :0.523
##                               3rd Qu.:38.00 3rd Qu.:1.000
##                               Max.   :80.00  Max.   :8.000
##                               NA's    :177
##   Parch      Ticket      Fare     Cabin Embarked
##   Min.   :0.0000  1601      : 7  Min.   : 0.00  B96   B98   : 4      : 0
##   1st Qu.:0.0000  347082    : 7  1st Qu.: 7.91  C23   C25   C27: 4  C    :168
##   Median :0.0000  CA. 2343: 7  Median :14.45  G6    : 4  Q    : 77
##   Mean   :0.3816  3101295 : 6  Mean   :32.20  C22   C26   : 3  S    :644
##   3rd Qu.:0.0000  347088  : 6  3rd Qu.:31.00  D    : 3  NA's: 2
##   Max.   :6.0000  CA 2144 : 6  Max.   :512.33 (Other) :186
##                   (Other) :852          NA's   :687
##   planta
##   C      : 59
##   B      : 47
##   D      : 33
##   E      : 32
##   A      : 15
##   (Other): 18
##   NA's   :687
```

Finalmente, vamos a dividir a los pasajeros en 3 grupos según su planta, planta alta, media y baja.

```
data$catplant <- factor(factor(data$planta, levels = c("A", "B", "C", "D", "E", "F", "F E", "F G", "G", "T"), labels = c("alta", "alta", "alta", "alta", "media", "baja", "baja", "baja", "baja"))
data$catplant <- relevel(data$catplant, ref = "baja")

summary(data)
```

```

##   Survived Pclass      Sex       Age      SibSp
## NO :549   1st:216 female:314   Min.   : 0.42   Min.   :0.000
## YES:342   2nd:184 male :577    1st Qu.:20.12  1st Qu.:0.000
##                   3rd:491          Median :28.00  Median :0.000
##                               Mean   :29.70  Mean   :0.523
##                               3rd Qu.:38.00 3rd Qu.:1.000
##                               Max.   :80.00  Max.   :8.000
##                               NA's   :177
##   Parch      Ticket     Fare      Cabin Embarked
## Min.   :0.0000  1601      : 7  Min.   : 0.00  B96 B98   : 4   : 0
## 1st Qu.:0.0000  347082    : 7  1st Qu.: 7.91  C23 C25 C27: 4   C   :168
## Median :0.0000  CA. 2343: 7  Median :14.45  G6      : 4   Q   : 77
## Mean   :0.3816  3101295: 6  Mean   :32.20  C22 C26   : 3   S   :644
## 3rd Qu.:0.0000  347088    : 6  3rd Qu.:31.00  D      : 3   NA's: 2
## Max.   :6.0000  CA 2144  : 6  Max.   :512.33 (Other) :186
##                   (Other) :852           NA's   :687
##   planta      catplant
##   C      : 59  baja  :18
##   B      : 47  alta  :154
##   D      : 33  media :32
##   E      : 32  NA's  :687
##   A      : 15
## (Other): 18
## NA's  :687

```

```
head(data, 5)
```

```

##   Survived Pclass      Sex Age SibSp Parch      Ticket     Fare Cabin
## 1      NO   3rd male  22   1   0     A/5 21171 7.2500 <NA>
## 2     YES   1st female 38   1   0     PC 17599 71.2833 C85
## 3     YES   3rd female 26   0   0     0 STON/O2. 3101282 7.9250 <NA>
## 4     YES   1st female 35   1   0     113803 53.1000 C123
## 5      NO   3rd male  35   0   0     373450 8.0500 <NA>
##   Embarked planta catplant
## 1      S <NA> <NA>
## 2      C   C alta
## 3      S <NA> <NA>
## 4      S   C alta
## 5      S <NA> <NA>

```

### 3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Ahora que tenemos todas las variables cargadas correctamente, podemos observar que tenemos valores faltantes en las variables **Age**, **Cabin** y **Embarked**, en la variable **Embarked** solo tenemos 2 valores faltantes, por lo tanto, vamos a eliminarlos, ya que representan una fracción muy pequeña del total de información de la muestra. En la variable **Age** tenemos un total de 117 valores faltantes, imputaremos los valores haciendo uso de la técnica missForest.

```

data <- data[!is.na(data$Embarked),]
dat_impo <- missForest(data[,c('Age', 'Sex', 'Fare')])

```

```

## missForest iteration 1 in progress...done!
## missForest iteration 2 in progress...done!
## missForest iteration 3 in progress...done!
## missForest iteration 4 in progress...done!

```

```

data$Age <- dat_impo$ximp$Age
summary(data)

```

```

##   Survived Pclass      Sex       Age      SibSp
## NO :549   1st:214   female:312   Min.   : 0.42   Min.   :0.0000
## YES:340   2nd:184   male   :577   1st Qu.:22.00   1st Qu.:0.0000
##                   3rd:491                               Median :29.57   Median :0.0000
##                                         Mean   :29.64   Mean   :0.5242
##                                         3rd Qu.:35.07   3rd Qu.:1.0000
##                                         Max.   :80.00   Max.   :8.0000
##
##      Parch      Ticket      Fare      Cabin Embarked
## Min.   :0.0000   1601   : 7   Min.   : 0.000   B96 B98   : 4   : 0
## 1st Qu.:0.0000   347082  : 7   1st Qu.: 7.896   C23 C25 C27: 4   C:168
## Median :0.0000   CA. 2343: 7   Median :14.454   G6          : 4   Q: 77
## Mean   :0.3825   3101295: 6   Mean   :32.097   C22 C26   : 3   S:644
## 3rd Qu.:0.0000   347088  : 6   3rd Qu.:31.000   D          : 3
## Max.   :6.0000   CA 2144 : 6   Max.   :512.329   (Other)    :184
##                   (Other) :850   NA's        :687
##
##      planta      catplant
## C      : 59   baja : 18
## B      : 45   alta :152
## D      : 33   media: 32
## E      : 32   NA's :687
## A      : 15
## (Other): 18
## NA's  :687

```

En cambio, la variable **Cabin**, que es, de las que podemos presuponer, más interesantes para el estudio, posee un total de 687 registros en los que su valor es NA, pero tenemos la variable **Pclass** y podemos pensar que la mayoría de habitaciones de una misma clase se encontrarían en una misma **planta**, vamos a intentar ver gráficamente si esta asunción es correcta.

```
nrow(data[data$Pclass == "1st" & data$catplant == "alta",])
```

```
## [1] 188
```

```
nrow(data[data$Pclass == "1st" & data$catplant == "media",])
```

```
## [1] 65
```

```
nrow(data[data$Pclass == "1st" & data$catplant == "baja",])
```

```
## [1] 41
```

```
nrow(data[data$Pclass == "2nd" & data$catplant == "alta",])
```

```
## [1] 172
```

```
nrow(data[data$Pclass == "2nd" & data$catplant == "media",])
```

```
## [1] 172
```

```
nrow(data[data$Pclass == "2nd" & data$catplant == "baja",])
```

```
## [1] 176
```

```
nrow(data[data$Pclass == "3rd" & data$catplant == "alta",])
```

```
## [1] 479
```

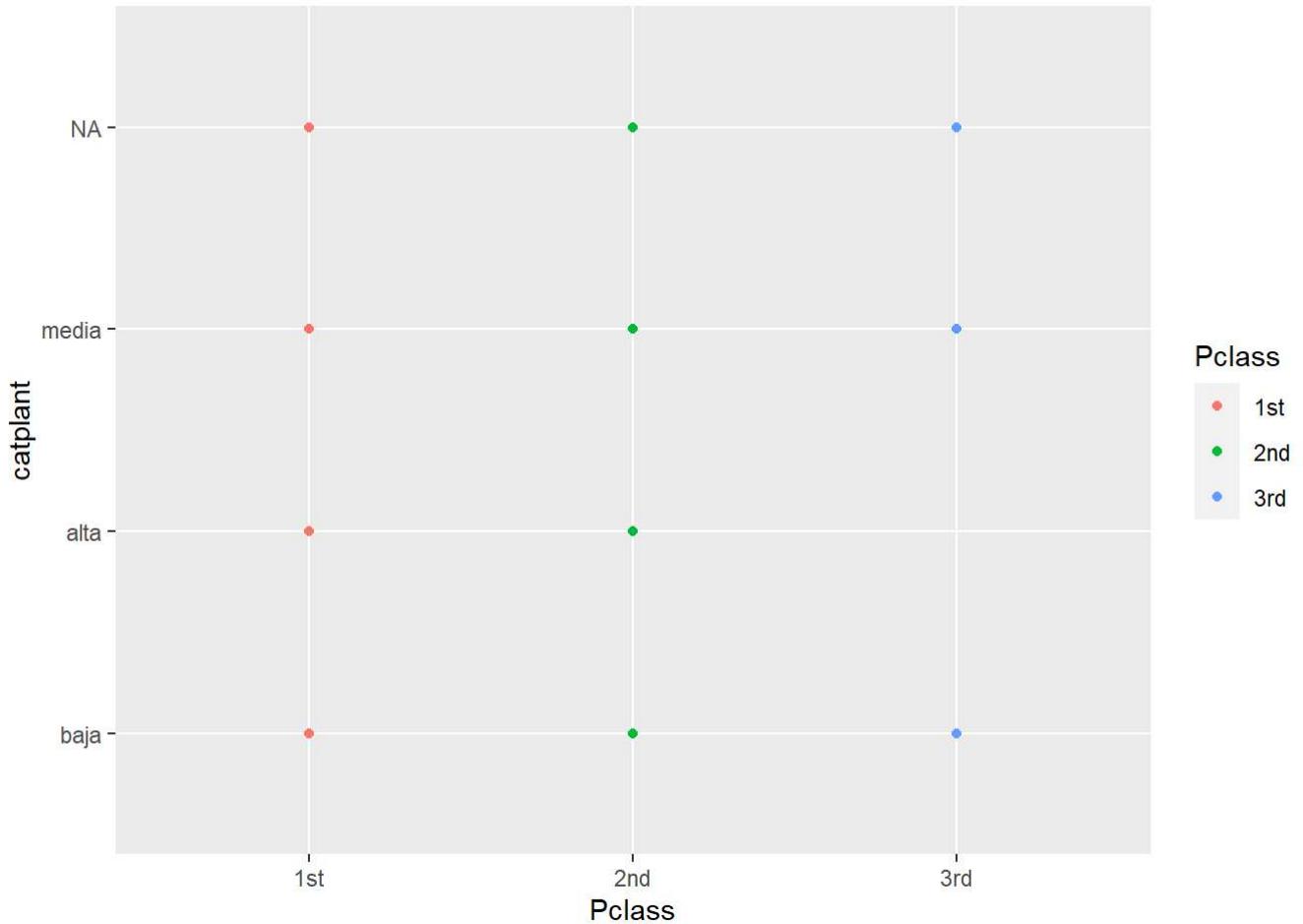
```
nrow(data[data$Pclass == "3rd" & data$catplant == "media",])
```

```
## [1] 482
```

```
nrow(data[data$Pclass == "3rd" & data$catplant == "baja",])
```

```
## [1] 488
```

```
qplot(Pclass, catplant, color = Pclass, data = data)
```



Podemos ver, mediante el conteo y el gráfico, que había pasajeros de todos los niveles en todas las alturas del barco, además, vemos que mientras que los pasajeros de primera si se situaban en su mayoría en las plantas más elevadas, los pasajeros de segunda y tercera, estaban repartidos entre todos los niveles, como no encontramos otra posible relación natural entre las otras variables y la variable **Cabin** vamos a generar una tabla de correlaciones a ver si observamos algún tipo de dependencia:

```

## Funcion para La matriz de correlacion
cor2 = function(df){

  stopifnotinherits(df, "data.frame"))
  stopifnot(sapply(df, class) %in% c("integer"
                                    , "numeric"
                                    , "factor"
                                    , "character"))

  cor_fun <- function(pos_1, pos_2){

    # both are numeric
    if(class(df[[pos_1]]) %in% c("integer", "numeric") &&
       class(df[[pos_2]]) %in% c("integer", "numeric")){
      r <- stats::cor(df[[pos_1]]
                      , df[[pos_2]]
                      , use = "pairwise.complete.obs"
      )
    }

    # one is numeric and other is a factor/character
    if(class(df[[pos_1]]) %in% c("integer", "numeric") &&
       class(df[[pos_2]]) %in% c("factor", "character")){
      r <- sqrt(
        summary(
          stats::lm(df[[pos_1]] ~ as.factor(df[[pos_2]])))[[["r.squared"]])
    }

    if(class(df[[pos_2]]) %in% c("integer", "numeric") &&
       class(df[[pos_1]]) %in% c("factor", "character")){
      r <- sqrt(
        summary(
          stats::lm(df[[pos_2]] ~ as.factor(df[[pos_1]])))[[["r.squared"]]))
    }

    # both are factor/character
    if(class(df[[pos_1]]) %in% c("factor", "character") &&
       class(df[[pos_2]]) %in% c("factor", "character")){
      r <- lsr::cramersV(df[[pos_1]], df[[pos_2]], simulate.p.value = TRUE)
    }

    return(r)
  }

  cor_fun <- Vectorize(cor_fun)

  # now compute corr matrix
  corrrmat <- outer(1:ncol(df)
                    , 1:ncol(df)
                    , function(x, y) cor_fun(x, y)
  )

  rownames(corrrmat) <- colnames(df)
  colnames(corrrmat) <- colnames(df)

  return(corrrmat)
}

```

```
#Generamos matriz de correlación
cor_matrix = cor2(data)
cor_matrix
```

```
##          Survived      Pclass       Sex        Age      SibSp      Parch
## Survived 1.00000000 0.33702932 0.5415849  0.08362164  0.03404000  0.08315078
## Pclass    0.33702932 1.00000000 0.1347433  0.35097449  0.09168361  0.01693923
## Sex      0.54158492 0.13474332 1.0000000  0.11121411  0.11634817  0.24750798
## Age      0.08362164 0.35097449 0.1112141  1.00000000 -0.21406434 -0.17427121
## SibSp    0.03404000 0.09168361 0.1163482 -0.21406434  1.00000000  0.41454164
## Parch    0.08315078 0.01693923 0.2475080 -0.17427121  0.41454164  1.00000000
## Ticket   0.92727635 1.00000000 0.8591533  0.88114033  0.94781362  0.90621667
## Fare     0.25529046 0.59309313 0.1799575  0.09959910  0.16088685  0.21753204
## Cabin    0.88848106 1.00000000 0.8573439  0.91741813  0.96388739  0.96977785
## Embarked 0.17261683 0.26382143 0.1225692  0.04764708  0.07004579  0.08503692
## planta   0.28725293 0.75951606 0.3233216  0.37097904  0.27695766  0.27153421
## catplant 0.10031552 0.54737907 0.0186515  0.30174715  0.09295963  0.11578545
##          Ticket      Fare      Cabin Embarked planta catplant
## Survived 0.9272763 0.2552905 0.8884811 0.17261683 0.2872529 0.10031552
## Pclass   1.0000000 0.5930931 1.0000000 0.26382143 0.7595161 0.54737907
## Sex     0.8591533 0.1799575 0.8573439 0.12256919 0.3233216 0.01865150
## Age     0.8811403 0.0995991 0.9174181 0.04764708 0.3709790 0.30174715
## SibSp   0.9478136 0.1608869 0.9638874 0.07004579 0.2769577 0.09295963
## Parch   0.9062167 0.2175320 0.9697779 0.08503692 0.2715342 0.11578545
## Ticket  1.0000000 1.0000000 0.9496240 0.99780859 0.9949447 0.99526015
## Fare    1.0000000 1.0000000 0.9409386 0.28155044 0.4459706 0.31993314
## Cabin   0.9496240 0.9409386 1.0000000 0.94926636 1.0000000 1.00000000
## Embarked 0.9978086 0.2815504 0.9492664 1.00000000 0.2583302 0.19857723
## planta  0.9949447 0.4459706 1.0000000 0.25833023 1.0000000 1.00000000
## catplant 0.9952601 0.3199331 1.0000000 0.19857723 1.0000000 1.00000000
```

Después de ver qué **Cabin** tiene una alta correlación con muchas de las otras variables, debemos plantearnos que los valores faltantes no se deban a un error, si no que sea un valor valido que represente a los tripulantes que viajaban sin una habitación asignada, después de investigar hemos constatado que el Titanic contaba con un total de 365 habitaciones, por lo que, es normal que no todos los pasajeros tuvieran una asignada, siendo, por tanto, NA un valor válido de esta variable.

## 3.2. Identificación y tratamiento de valores extremos.

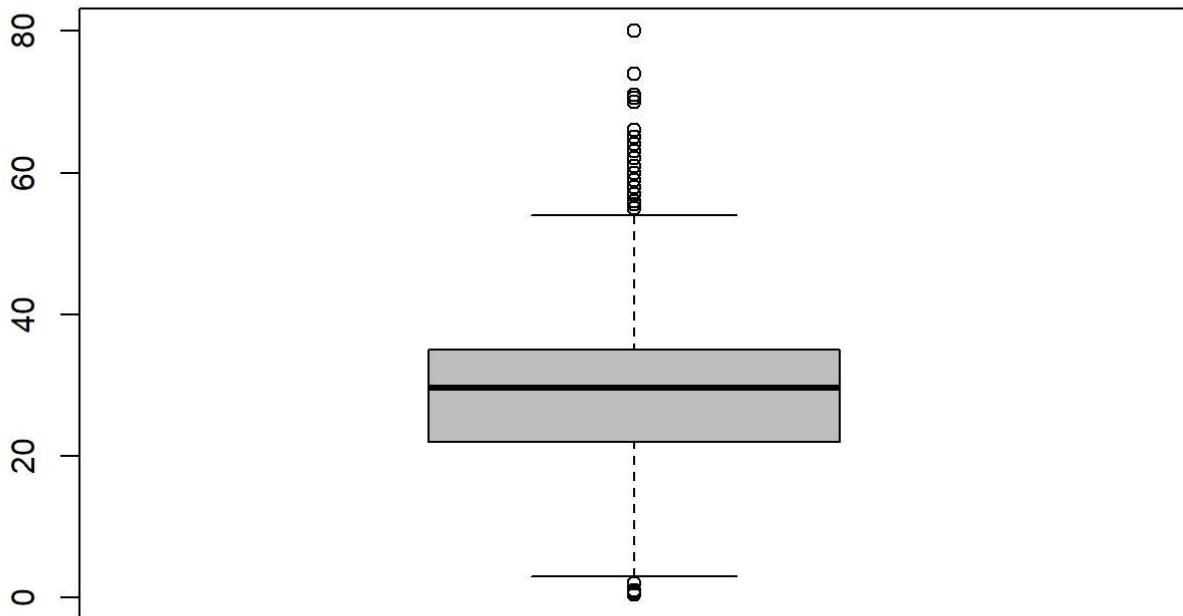
A continuación, mediante diagramas de cajas, vamos a buscar valores extremos en nuestras variables numéricas. Empezaremos por la variable **Age**:

```
length(data$Age)
```

```
## [1] 889
```

```
boxplot(data$Age,main="Age", col="gray")
```

## Age



```
extremos_Age <- boxplot.stats(data$Age)$out  
extremos_Age
```

```
## [1] 2.00 58.00 55.00 2.00 66.00 65.00 0.83 59.00 71.00 70.50 2.00 55.50  
## [13] 1.00 61.00 1.00 56.00 1.00 58.00 2.00 59.00 62.00 58.00 63.00 65.00  
## [25] 2.00 0.92 61.00 2.00 60.00 1.00 1.00 64.00 65.00 56.00 0.75 2.00  
## [37] 63.00 58.00 55.00 71.00 2.00 64.00 62.00 62.00 60.00 61.00 57.00 80.00  
## [49] 2.00 0.75 56.00 58.00 70.00 60.00 60.00 70.00 0.67 57.00 1.00 0.42  
## [61] 2.00 1.00 0.83 74.00 56.00
```

```
length(extremos_Age)
```

```
## [1] 65
```

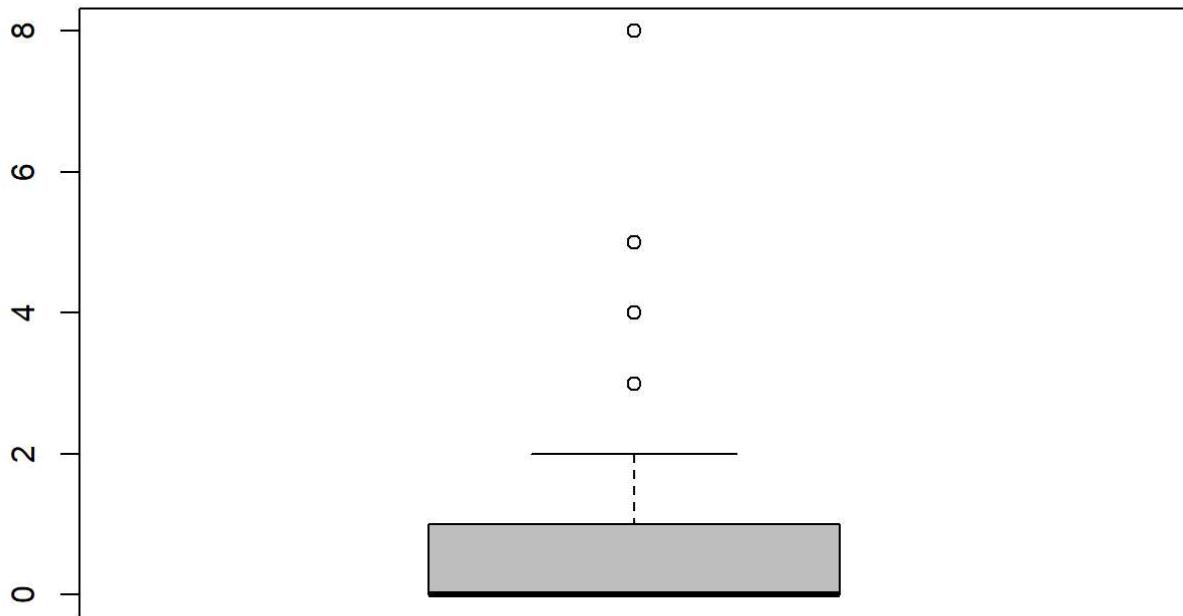
Podemos observar que detectamos 8 valores como valores extremos, pero como representan una parte muy pequeña del total de la muestra, con bajo riesgo para sesgar nuestros análisis y se encuentran dentro del dominio, no vamos a tratar estos registros. A continuación, vamos a observar la variable **SibSp**:

```
length(data$SibSp)
```

```
## [1] 889
```

```
boxplot(data$SibSp, main="SibSp", col="gray")
```

## SibSp



```
extremos_SibSp <- boxplot.stats(data$SibSp)$out  
extremos_SibSp
```

```
## [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
```

```
length(extremos_SibSp)
```

```
## [1] 46
```

En este caso tenemos un valor más significativo de valores extremos 46 de un total de 889 registros, pero de nuevo se encuentran dentro de lo que podríamos considerar un dominio aceptable, por lo que no vamos a considerarlos outliers. Finalmente vamos a revisar la variable **Parch**:

```
length(data$Parch)
```

```
## [1] 889
```

```
boxplot(data$Parch, main="Parch", col="gray")
```



```
extremos_Parch <- boxplot.stats(data$Parch)$out  
extremos_Parch
```

length(extremos Parch)

```
## [1] 213
```

De nuevo, volvemos a tener más valores extremos, pero siguen dentro del rango de la variable y no suponen un riesgo real de sesgar nuestra muestra, los tendremos especialmente en cuenta, pero no vamos a considerarlos outliers.

# 4. Análisis de los datos.

## 4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Vamos a analizar los datos utilizando las variables **Pclass**, **catplant**, **Sex**, **Fare** y **Age**

A priori, nos parecen variables representativas del dataset cara a hacer distintos análisis.

De forma informal siempre que hay un naufragio se dice que las mujeres y los niños primero. Nos gustaría ver si realmente fue así en el Titanic, por lo que realizaremos un análisis al respecto.

También queremos ver si la edad puede influir en el hecho de haber sobrevivido o no, nos gustaría analizar si se salvaron más jóvenes en media que personas mayores.

Haremos distintos análisis focalizándonos especialmente en sexo y edad mediante hipótesis y regresiones, además incluiremos un arbol de decisión con estas variables ver qué tipos de reglas podemos extraer.

Por último, incluimos una correlación para ver cómo se relaciona el precio del ticket (**Fare**) con la edad para ver si el precio del ticket comprado crece o decrece con la edad.

## 4.2. Comprobación de la normalidad y homogeneidad de la varianza.

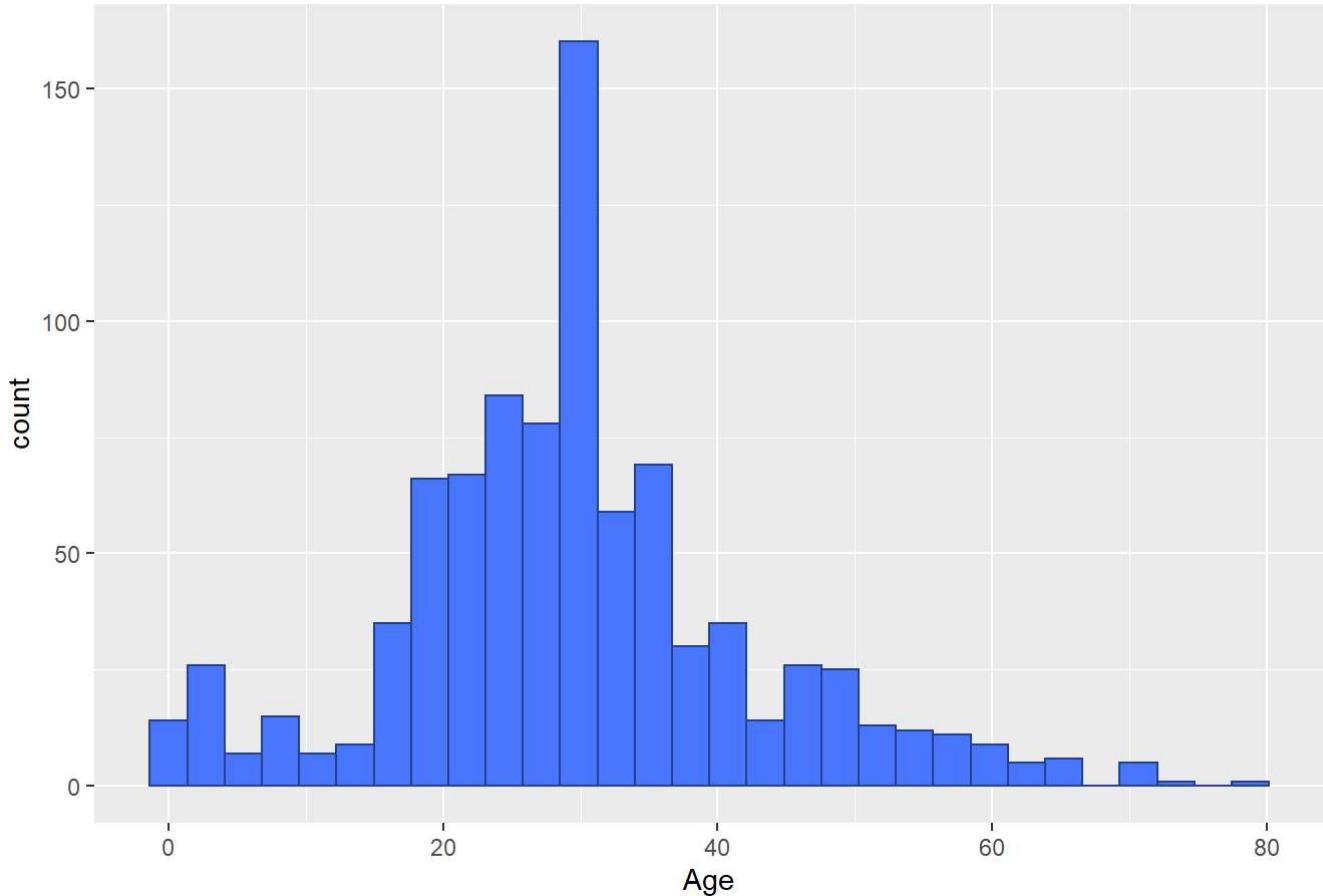
A continuación, vamos a comprobar el supuesto de normalidad, para un nivel de confianza del 95% sobre la variable **Age** y **Fare**, ya que son las únicas de las variables mencionadas anteriormente que son continuas, mediante el test de Shapiro-Wilk:

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

```
##  
##  Shapiro-Wilk normality test  
##  
## data: data$Age  
## W = 0.96876, p-value = 7.12e-13
```

Podemos ver que el p-valor devuelto por el test es 5.97e-13, por tanto, tomaremos como válida la hipótesis alternativa del test, considerando que la variable no sigue una distribución normal. Podemos observarlo gráficamente:

```
rows = nrow(data)  
bw <- 2 * IQR(data$Age) / length(data$Age)^(1/3)  
Age_g <- ggplot(data = data[1:rows,], aes(Age))  
Age_g <- Age_g + geom_histogram(col="royalblue4", fill="royalblue1", binwidth = bw)  
grid.arrange(Age_g, nrow = 1, ncol = 1, top = " ")
```



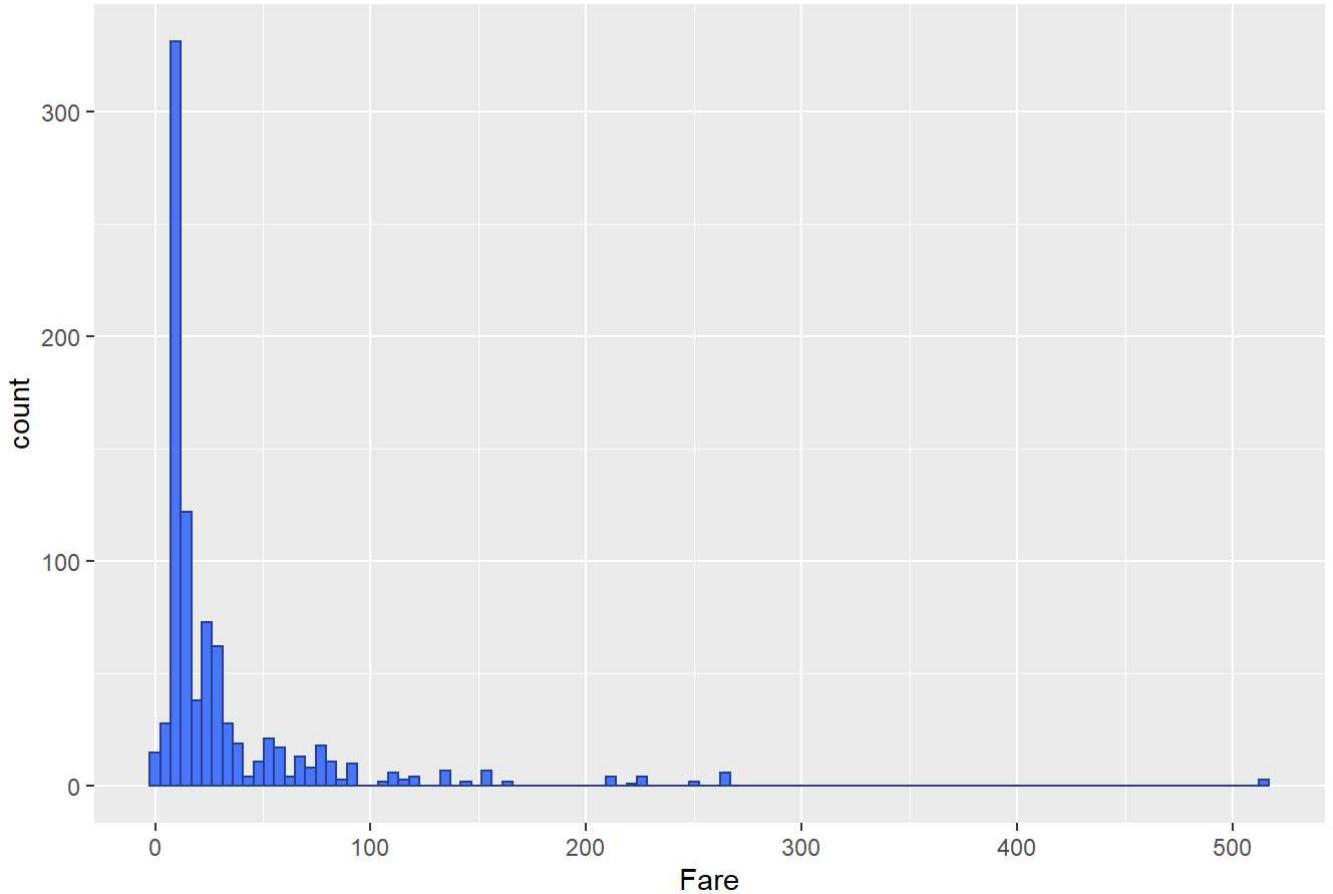
Podemos concluir, por tanto, que realmente la variable **Age** no sigue una distribución normal. A continuación vamos a realizar este mismo test para la variable **Fare**.

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

```
##  
## Shapiro-Wilk normality test  
##  
## data: data$Fare  
## W = 0.5197, p-value < 2.2e-16
```

Podemos ver que el p-valor devuelto por el test es 2.2e-16, por tanto, tomaremos como válida la hipótesis alternativa del test, considerando que la variable no sigue una distribución normal. Podemos observarlo gráficamente:

```
rows = nrow(data)  
bw <- 2 * IQR(data$Fare) / length(data$Fare)^(1/3)  
Age_g <- ggplot(data = data[1:rows,], aes(Fare))  
Age_g <- Age_g + geom_histogram(col="royalblue4", fill="royalblue1", binwidth = bw)  
grid.arrange(Age_g, nrow = 1, ncol = 1, top = " ")
```



En cuanto al test de homocedasticidad, vamos a realizarlo para nuestras 3 variables categóricas y para el cruce de **Age** y **Fare** con la variable respuesta **Survived**, vamos a hacer uso del test de Fligner-Killeen dado que es el más robusto para variables que se alejan de una distribución normal:

```
fligner.test(as.numeric(Survived) ~ Pclass, data = data)
```

```
## 
##  Fligner-Killeen test of homogeneity of variances
## 
##  data:  as.numeric(Survived) by Pclass
##  Fligner-Killeen:med chi-squared = 36.046, df = 2, p-value = 1.488e-08
```

```
fligner.test(as.numeric(Survived) ~ catplant, data = data)
```

```
## 
##  Fligner-Killeen test of homogeneity of variances
## 
##  data:  as.numeric(Survived) by catplant
##  Fligner-Killeen:med chi-squared = 2.0227, df = 2, p-value = 0.3637
```

```
fligner.test(as.numeric(Survived) ~ Sex, data = data)
```

```
##  
## Fligner-Killeen test of homogeneity of variances  
##  
## data: as.numeric(Survived) by Sex  
## Fligner-Killeen:med chi-squared = 6.0178, df = 1, p-value = 0.01416
```

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

```
##  
## Fligner-Killeen test of homogeneity of variances  
##  
## data: as.numeric(Survived) by Age  
## Fligner-Killeen:med chi-squared = 170.43, df = 161, p-value = 0.2903
```

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

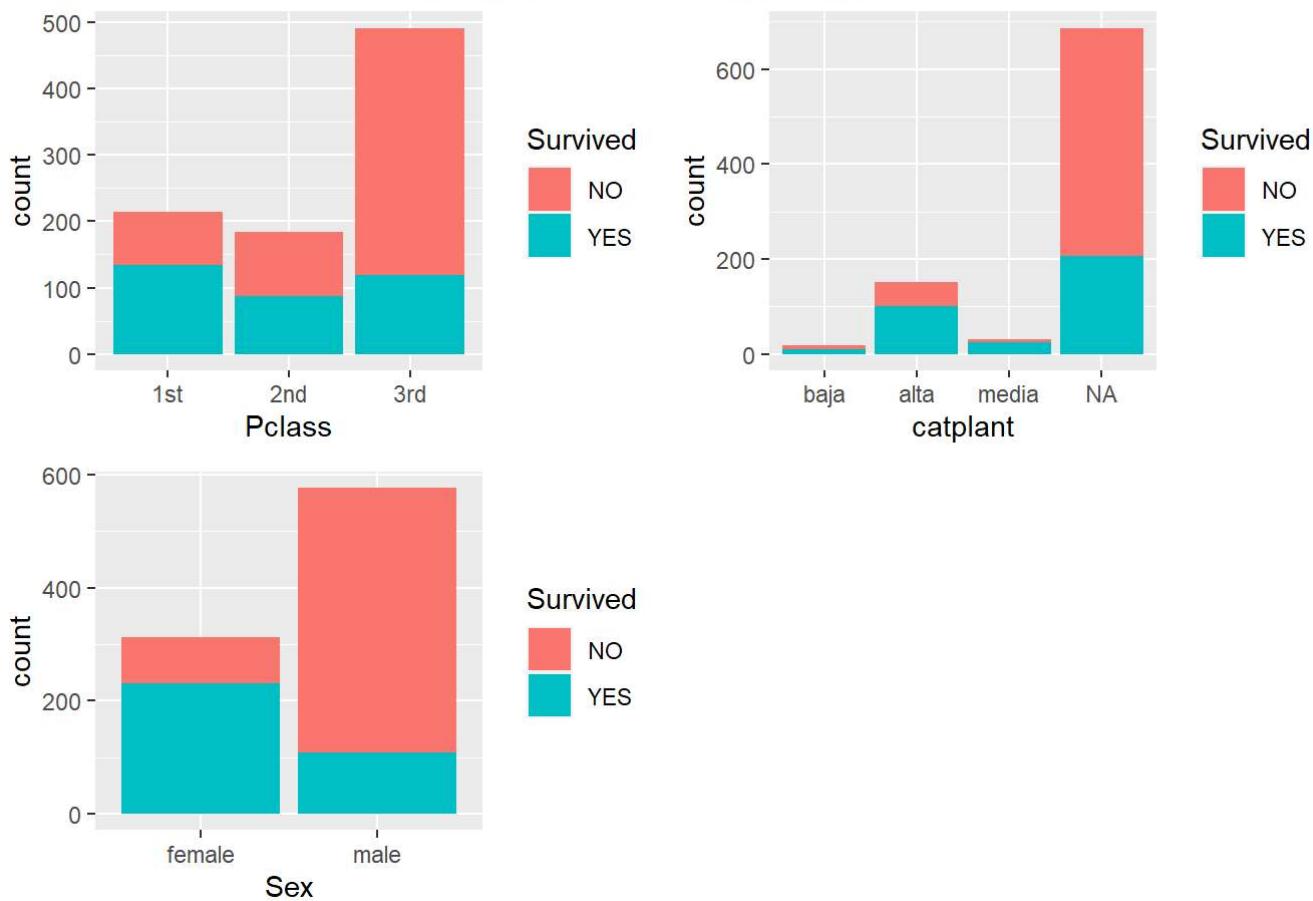
```
##  
## Fligner-Killeen test of homogeneity of variances  
##  
## data: as.numeric(Survived) by Fare  
## Fligner-Killeen:med chi-squared = 257.1, df = 246, p-value = 0.3004
```

Vemos que hay una varianza no homogénea solo en la variable **Pclass**, lo cual se ve reforzado por la correlación de esta variable con la variable **Survived**, que es de 0.33702932, en cambio vemos que las otras dos variables que si presentan mayor varianza, como nos indica su correlación, de 0.10031552 para la variable **catplant**, de 0.54158492 para la variable **Sex**, de 0.07451322 para la variable **Age** y de 0.25529046 para la variable **Fare**.

Finalmente, vamos a visualizar la distribución de nuestras variables categóricas:

```
rows = dim(data)[1]  
  
Pclass_g <- ggplot(data = data[1:rows,], aes(x=Pclass, fill=Survived))  
Pclass_g <- Pclass_g + geom_bar()  
  
catplant_g <- ggplot(data = data[1:rows,], aes(x=catplant, fill=Survived))  
catplant_g <- catplant_g + geom_bar()  
  
Sex_g <- ggplot(data = data[1:rows,], aes(x=Sex, fill=Survived))  
Sex_g <- Sex_g + geom_bar()  
  
grid.arrange(Pclass_g, catplant_g, Sex_g, nrow = 2, ncol = 2, top = "Distribución de las variables categoricas")
```

Distribución de las variables categoricas



## 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

### 4.3.1 Hipótesis edad supervivientes

Una pregunta que nos podemos hacer después de haber revisado las edades de los pasajeros, es si los supervivientes del Titanic eran más jóvenes que quienes no sobrevivieron, con un rango de confianza del 95%.

De este modo, podemos lanzar la siguiente hipótesis nula y su alternativa:

- H<sub>0</sub>: La media de edad de los supervivientes del Titanic = que media de edad de los que no sobrevivieron
- H<sub>1</sub>: La media de edad de los supervivientes del Titanic < que media de edad de los que no sobrevivieron

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

Ejecutamos un t-tes para hacer el contrate de hipótesis:

```
t.test(Age ~ Survived, data)
```

```
##  
## Welch Two Sample t-test  
##  
## data: Age by Survived  
## t = 2.4445, df = 667.28, p-value = 0.01476  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.4408151 4.0401245  
## sample estimates:  
## mean in group NO mean in group YES  
## 30.49849 28.25802
```

Vemos que p-value es menor que 0.05, por lo que podemos rechazar la hipótesis nula a favor de la hipótesis alternativa, por lo que podemos decir que la media de edad de los que sobrevivieron era menor.

De hecho, la media en el grupo de supervivientes es de 28 años y en el grupo de no supervivientes es de 30 años.

### 4.3.2 Hipótesis sexo supervivientes

Podemos hacer lo mismo a nivel de género de los supervivientes.

Utilizamos la función **crosstable** para ver las proporciones de supervivientes y muertos según género.

```
CrossTable(data$Survived,data$Sex)
```

```

##  

##  

##      Cell Contents  

## |-----|  

## |           N |  

## | Chi-square contribution |  

## |           N / Row Total |  

## |           N / Col Total |  

## |           N / Table Total |  

## |-----|  

##  

##  

## Total Observations in Table: 889  

##  

##  

##          | data$Sex  

## data$Survived | female | male | Row Total |  

## -----|-----|-----|-----|  

## NO |     81 |   468 |    549 |  

## | 64.727 | 35.000 |  

## | 0.148 | 0.852 | 0.618 |  

## | 0.260 | 0.811 |  

## | 0.091 | 0.526 |  

## -----|-----|-----|-----|  

## YES |   231 |   109 |    340 |  

## | 104.515 | 56.514 |  

## | 0.679 | 0.321 | 0.382 |  

## | 0.740 | 0.189 |  

## | 0.260 | 0.123 |  

## -----|-----|-----|-----|  

## Column Total | 312 | 577 | 889 |  

## | 0.351 | 0.649 |  

## -----|-----|-----|-----|
##  

##
```

De la tabla anterior, podemos decir que el 67.9% de los supervivientes eran mujeres y que el 74% de las mujeres que estaban a bordo sobrevivieron.

Podemos pensar que hay una relación entre género y supervivencia, por lo que podemos formular la siguiente hipótesis.

H0 - No hay relación entre ambas variables

H1 - Hay una relación entre ambas variables

Para lanzar esta hipótesis, podemos utilizar el test Chi-cuadrado ya son datos categóricos.

```

a <- xtabs(~Survived+Sex, data)
chisq.test(a)
```

```

##  

## Pearson's Chi-squared test with Yates' continuity correction  

##  

## data: a  

## X-squared = 258.43, df = 1, p-value < 2.2e-16
```

Como podemos ver, el **p-value** es mucho menor que 0.05, por lo que podemos rechazar la hipótesis nula. Por tanto ambas variables de género y supervivencia están relacionadas.

### 4.3.2 Regresión con edad

Podemos analizar si la variable **Survived** que indica si el pasajero sobrevivió es una variable dependiente de la edad y su grado de dependencia.

Podemos utilizar una regresión logística simple para analizar este punto:

```
model <- glm(Survived ~ Age, data, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ Age, family = binomial, data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.1362 -0.9890 -0.9321  1.3544  1.6928
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.084198   0.172122 -0.489   0.6247
## Age         -0.013450   0.005418 -2.482   0.0131 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1182.8 on 888 degrees of freedom
## Residual deviance: 1176.5 on 887 degrees of freedom
## AIC: 1180.5
##
## Number of Fisher Scoring iterations: 4
```

A partir de los resultados ( $p=0.0268$ ), podemos decir que existe una relación lineal entre ambas variables pero su relación es baja.

A partir de esta regresión podríamos lanzar una pregunta del tipo, qué probabilidad tendría de sobrevivir en caso de que tuviera 10 años, 25 años y 74 años.

```
pred<-predict(model, data.frame(Age=10),type = "response")
pred
```

```
##      1
## 0.4455427
```

La probabilidad de sobrevivir con una edad de 10 años es de un 43.88%.

```
pred<-predict(model, data.frame(Age=25),type = "response")
pred
```

```
##      1  
## 0.3964102
```

La probabilidad de sobrevivir con una edad de 25 años sería de un 39.5%.

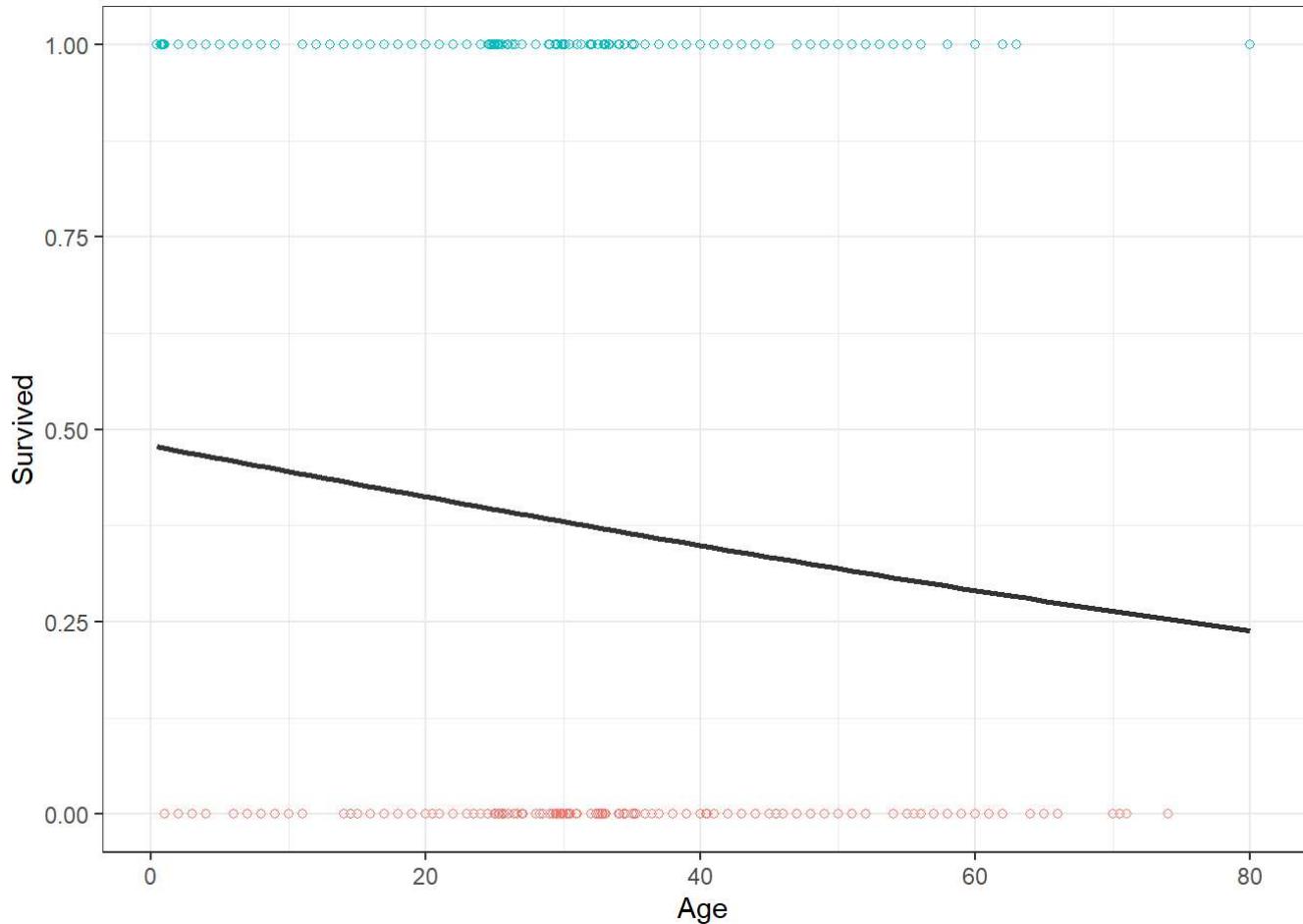
```
pred<-predict(model, data.frame(Age=74),type = "response")  
pred
```

```
##      1  
## 0.2536024
```

La probabilidad de sobrevivir con una edad de 74 años sería de un 26.6%.

A continuación graficamos la regresión. Podemos ver gráficamente que por ejemplo la probabilidad anterior de sobrevivir con 74 años es aproximadamente un 26%. Lo que podemos ver es que en base a esta regresión logística, conforme se va aumentando la edad, baja la probabilidad de supervivencia desde un 48% a un 25% aproximadamente.

```
# Primero tenemos que recodificar Survived a 1 y 0  
data_regresion <- data %>%  
  mutate(Survived = recode(Survived,  
    "NO" = 0,  
    "YES" = 1))  
  
# Procedemos a graficar la regresión  
ggplot(data_regresion, aes(x = Age, y = Survived)) +  
  geom_point(aes(color = as.factor(Survived)), shape = 1) +  
  geom_smooth(method = "glm",  
    method.args = list(family = "binomial"),  
    color = "gray20",  
    se = FALSE) +  
  theme_bw() +  
  theme(legend.position = "none")  
  
## `geom_smooth()` using formula 'y ~ x'
```

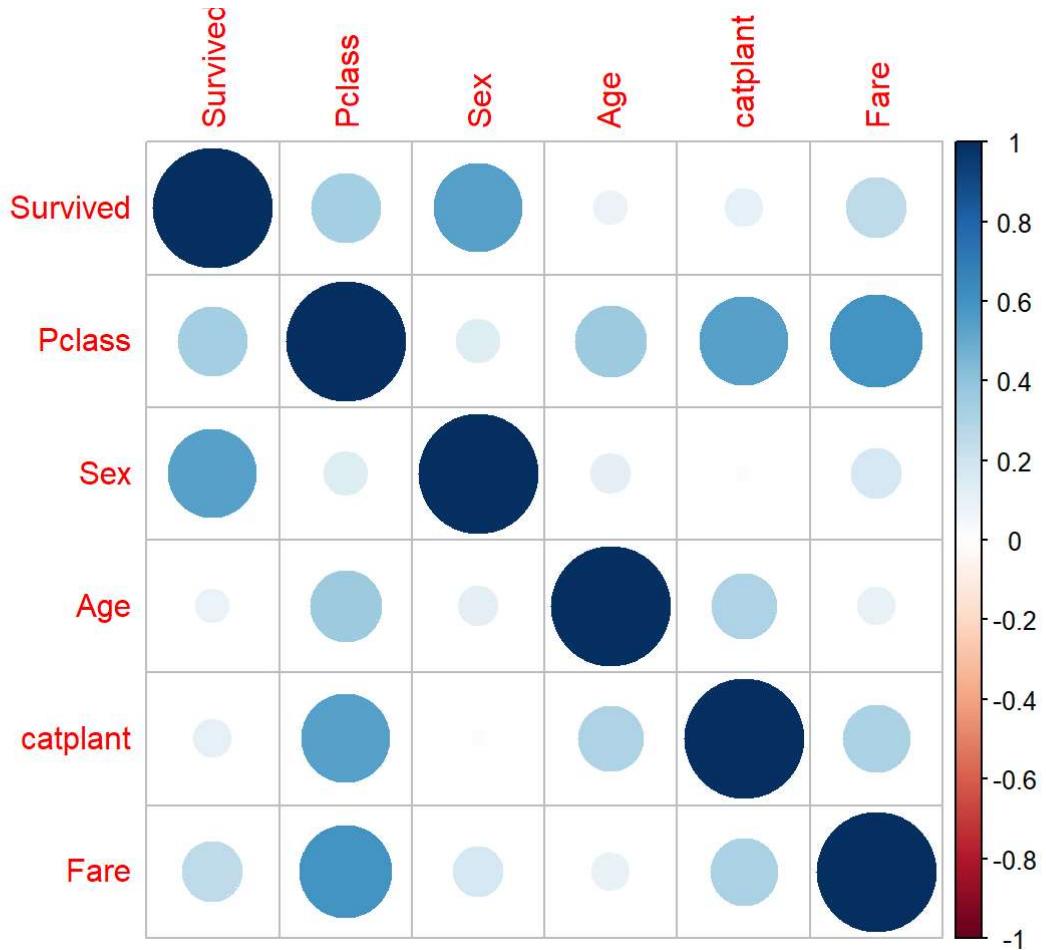


### 4.3.3 Correlación

También podemos revisar la correlación entre las distintas variables, podemos utilizar la función **corrplot** utilizando la función de generación de matrices de correlación que utilizamos al principio de este análisis. Lo aplicaremos sobre el grupo de variables que hemos seleccionado para esta sección. Aunque únicamente compararemos variables numéricas, por lo que podemos revisar la correlación entre **Age**, edad del pasajero, y **Fare**, precio de su billete.

```
data_selected <- select(data, Survived, Pclass, Sex, Age, catplant, Fare)
cor_matrix_selected = cor2(data_selected)

# Grafico la matriz de correlación
corrplot(cor_matrix_selected)
```



```
cor_matrix_selected
```

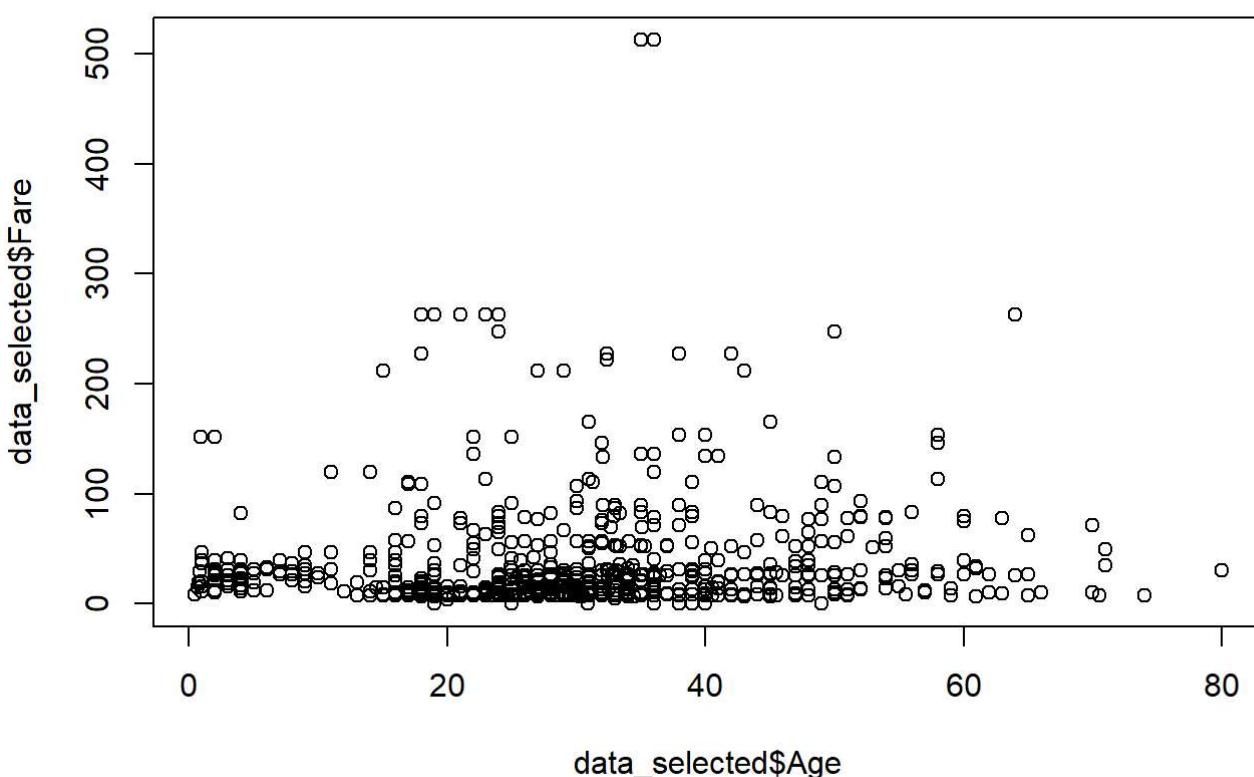
```
##           Survived      Pclass       Sex        Age     catplant      Fare
## Survived 1.00000000 0.3370293 0.5415849 0.08362164 0.1003155 0.2552905
## Pclass    0.33702932 1.0000000 0.1347433 0.35097449 0.5473791 0.5930931
## Sex      0.54158492 0.1347433 1.0000000 0.11121411 0.0186515 0.1799575
## Age     0.08362164 0.3509745 0.1112141 1.00000000 0.3017472 0.0995991
## catplant 0.10031552 0.5473791 0.0186515 0.30174715 1.0000000 0.3199331
## Fare    0.25529046 0.5930931 0.1799575 0.09959910 0.3199331 1.0000000
```

Si revisamos el coste del ticket (**Fare**) con respecto a la edad, tenemos una correlación de 0.10058049, por lo que relación entre estas variables sería muy baja.

Procedemos a graficarlo, se puede ver que no se sigue una progresión en el que al aumentar la edad o disminuir también aumente o disminuya el precio del ticket comprado.

```
plot(data_selected$Age, data_selected$Fare, main="Precio ticket vs Edad")
```

### Precio ticket vs Edad



La dispersión de este gráfico nos indica, del mismo modo que la correlación, que las variables **Age** y **Fare** no están relacionadas entre sí.

#### 4.3.4 Árbol de decisión

Vamos a utilizar un árbol de decisión para poder hacer predicción en base a estos datos. (**Pclass**, **catplant**, **Sex**, **Age** y **Fare**).

Nuestro objetivo es crear un árbol de decisión que permita analizar qué tipo de pasajero del Titanic tenía probabilidades de sobrevivir o no. Por lo tanto, la variable por la que clasificaremos es **Survived**.

En primer lugar nos vamos a quedar únicamente con los campos objeto del estudio y la clase que debemos predecir. (**Survived**).

```
data_selected <- select(data, Survived, Pclass, Sex, Age, catplant, Fare)

# Revisamos la estructura
str(data_selected)
```

```
## 'data.frame': 889 obs. of 6 variables:
## $ Survived: Factor w/ 2 levels "NO","YES": 1 2 2 2 1 1 1 1 2 2 ...
## $ Pclass   : Factor w/ 3 levels "1st","2nd","3rd": 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 ...
## $ catplant: Factor w/ 3 levels "baja","alta",...: NA 2 NA 2 NA NA 3 NA NA NA ...
## $ Fare     : num  7.25 71.28 7.92 53.1 8.05 ...
```

```
head(data_selected)
```

```

##   Survived Pclass     Sex     Age catplant     Fare
## 1      NO   3rd male 22.00000    <NA> 7.2500
## 2     YES   1st female 38.00000 alta 71.2833
## 3     YES   3rd female 26.00000    <NA> 7.9250
## 4     YES   1st female 35.00000 alta 53.1000
## 5      NO   3rd male 35.00000    <NA> 8.0500
## 6      NO   3rd male 29.19587    <NA> 8.4583

```

Vamos a desordenar un poco las filas para tener más aleatoriedad al crear el dataset de entrenamiento y el dataset de evaluación. El nuevo dataset desordenado lo almacenaremos en la variable **data\_random**.

```

set.seed(1)
data_random <- data_selected[sample(nrow(data_selected)),]

```

Creamos el conjunto de entrenamiento y el de evaluación, 2/3 de filas para el conjunto de entrenamiento y 1/3 para el conjunto de prueba.

```

set.seed(666)
y <- data_random[,1]
X <- data_random[,2:5]

indexes = sample(1:nrow(data_selected), size=floor((2/3)*nrow(data_selected)))
trainX<-X[indexes,]
trainy<-y[indexes]
testX<-X[-indexes,]
testy<-y[-indexes]

```

Se crea el árbol de decisión usando los datos de entrenamiento:

```

#model <- C50::C5.0(trainX, trainy, rules=TRUE )
#summary(model)

model_tree <- rpart(formula = trainy ~ ., data = cbind(trainX, trainy),
                     control = rpart.control(cp = 0, maxdepth = 4))
model_tree

```

```

## n= 592
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 592 226 NO (0.61824324 0.38175676)
##    2) Sex=male 386 74 NO (0.80829016 0.19170984)
##      4) Age>=6.5 370 61 NO (0.83513514 0.16486486) *
##      5) Age< 6.5 16 3 YES (0.18750000 0.81250000) *
##    3) Sex=female 206 54 YES (0.26213592 0.73786408)
##      6) Pclass=3rd 92 45 NO (0.51086957 0.48913043)
##        12) Age>=38.5 8 0 NO (1.00000000 0.00000000) *
##        13) Age< 38.5 84 39 YES (0.46428571 0.53571429)
##          26) Age< 21.5 34 14 NO (0.58823529 0.41176471) *
##          27) Age>=21.5 50 19 YES (0.38000000 0.62000000) *
##    7) Pclass=1st,2nd 114 7 YES (0.06140351 0.93859649) *

```

Del árbol podemos llegar a las siguientes conclusiones:

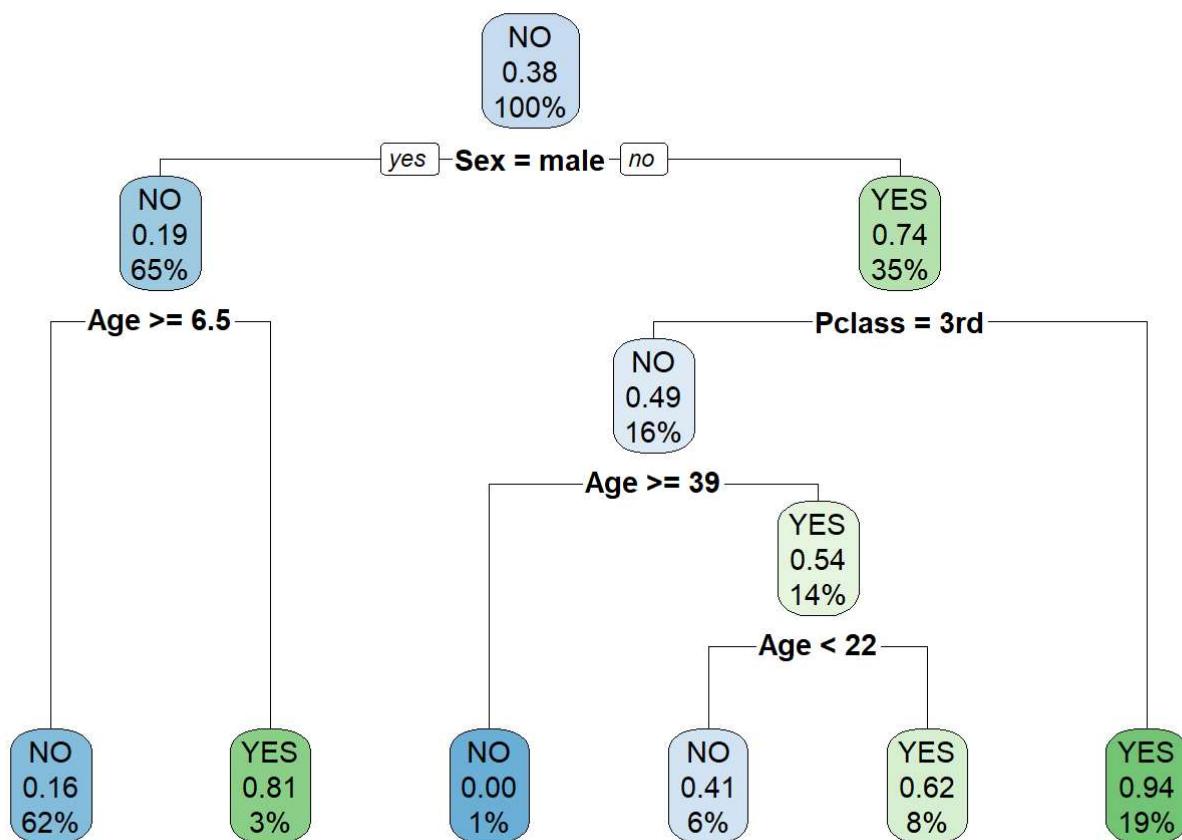
- Si es un hombre y tiene una edad  $\geq 6.5$  años, entonces muere con una probabilidad de un 80,83%
- Si es un hombre y tiene una edad  $< 6.5$  años, entonces sobrevive con una probabilidad del 81,25%
- Si es mujer
  - no viaja en tercera clase, entonces sobrevive con una probabilidad del 93,86%
  - viaja en tercera clase
    - su edad es mayor o igual que 38,5 años, entonces muere (100%)
    - su edad es menor de 21,5 años, entonces no sobrevive al 58,82%
    - su edad es menor de 38,5 años pero igual o mayor a 21,5 años, entonces sobrevive al 62%

Por tanto podemos concluir que el conocimiento extraído y cruzado con el análisis visual se resume en “las mujeres y los niños primero a excepción de que fueras de 3<sup>a</sup> clase”. Si eras de 3<sup>a</sup> clase tenías más posibilidades si eras mujer con una edad comprendida entre 21,5 y 38,5 años.

A continuación mostramos el árbol obtenido.

```
rpart.plot(model_tree)
```

```
## Warning: Bad 'data' field in model 'call' (expected a data.frame or a matrix).
## To silence this warning:
##   Call rpart.plot with roundint=FALSE,
##   or rebuild the rpart model with model=TRUE.
```



Una vez tenemos el modelo, podemos comprobar su calidad prediciendo la clase para los datos de prueba que nos hemos reservado al principio.

```
predicted_model <- predict( model_tree, testX, type="class" )
print(sprintf("La precisión del árbol es: %.4f %%", 100*sum(predicted_model == testy) / length(predicted_model)))
```

```
## [1] "La precisión del árbol es: 77.4411 %"
```

Tenemos a nuestra disposición el paquete **gmodels** para obtener información más completa:

```
CrossTable(testy, predicted_model, prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('Reality', 'Prediction'))
```

```
##
##
##      Cell Contents
## |-----|
## |           N |
## |           N / Table Total |
## |-----|
## 
## 
## Total Observations in Table:  297
##
##
##          | Prediction
##    Reality |      NO |      YES | Row Total |
## -----|-----|-----|-----|
##       NO |    160 |     23 |    183 |
##           | 0.539 | 0.077 |      |
## -----|-----|-----|-----|
##      YES |     44 |     70 |    114 |
##           | 0.148 | 0.236 |      |
## -----|-----|-----|-----|
## Column Total |    204 |     93 |    297 |
## -----|-----|-----|-----|
## 
```

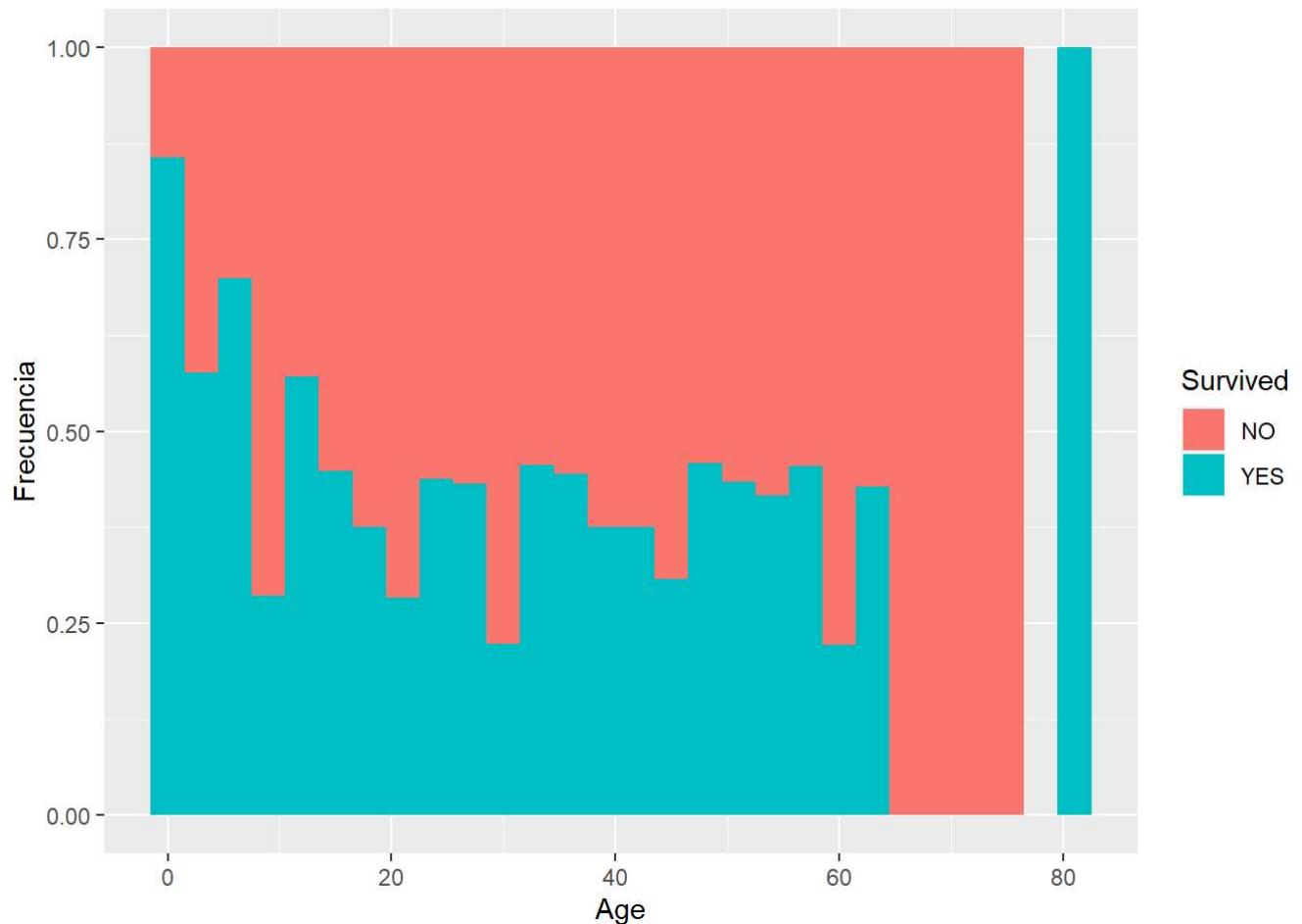
# 5. Representación gráfica

A continuación, hacemos una representación gráfica para visualizar las distintas conclusiones que hemos ido sacando de los análisis.

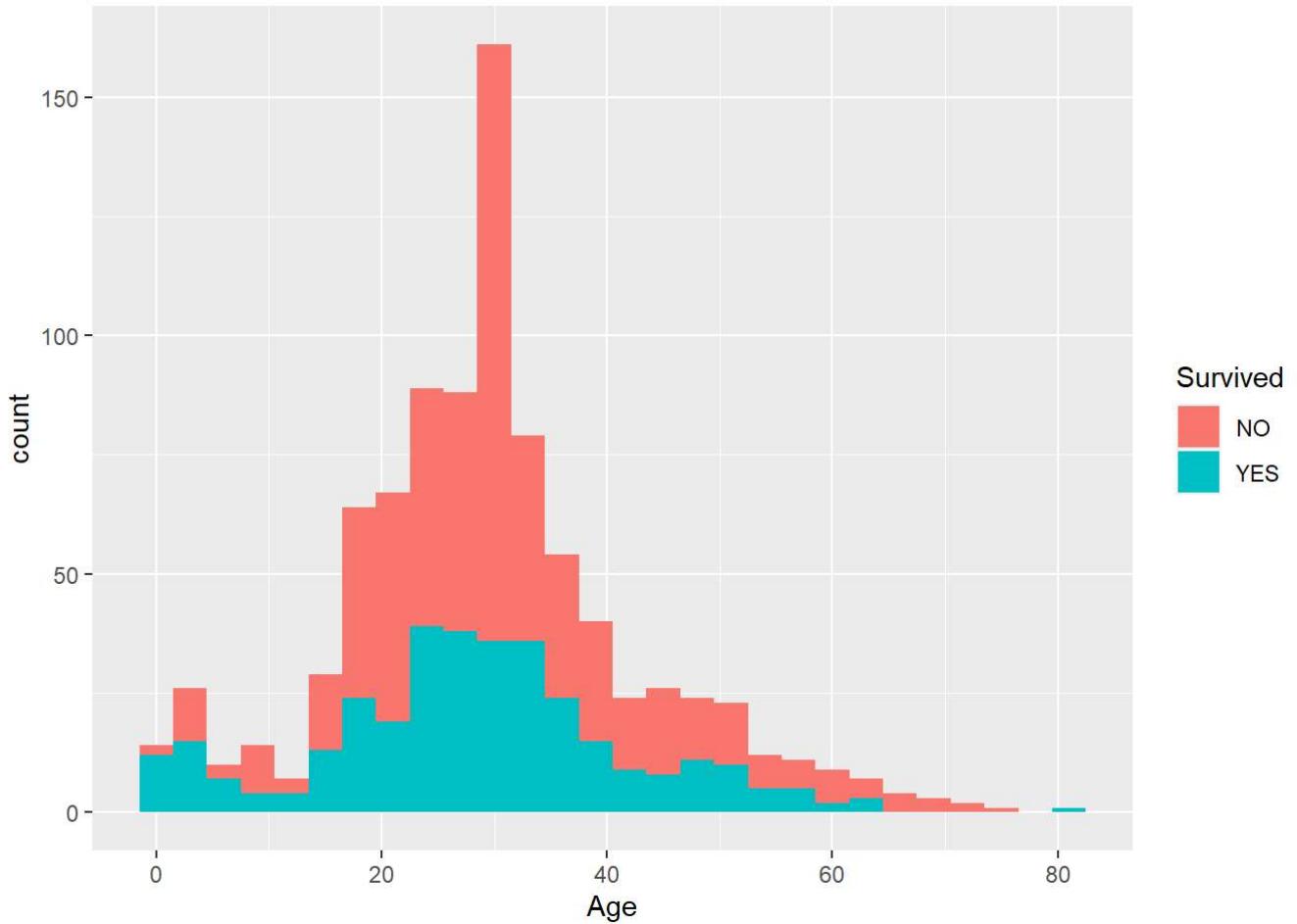
Como decíamos, parece que la edad y el sexo influye en la supervivencia, vamos a verlo gráficamente.

- Análisis por edad: podemos comprobar que a partir de 63 años aproximadamente, las opciones de supervivencia son de 0. Podemos ver que todos los mayores de 80 años sobrevivieron pero tan solo había una persona con esa edad. También se puede ver que quienes tenían mayores posibilidades de supervivencia eran los niños más pequeños.

```
ggplot(data_selected,aes(x=Age,fill=Survived))+geom_histogram(binwidth = 3,position="fill")+
  lab("Frecuencia")
```

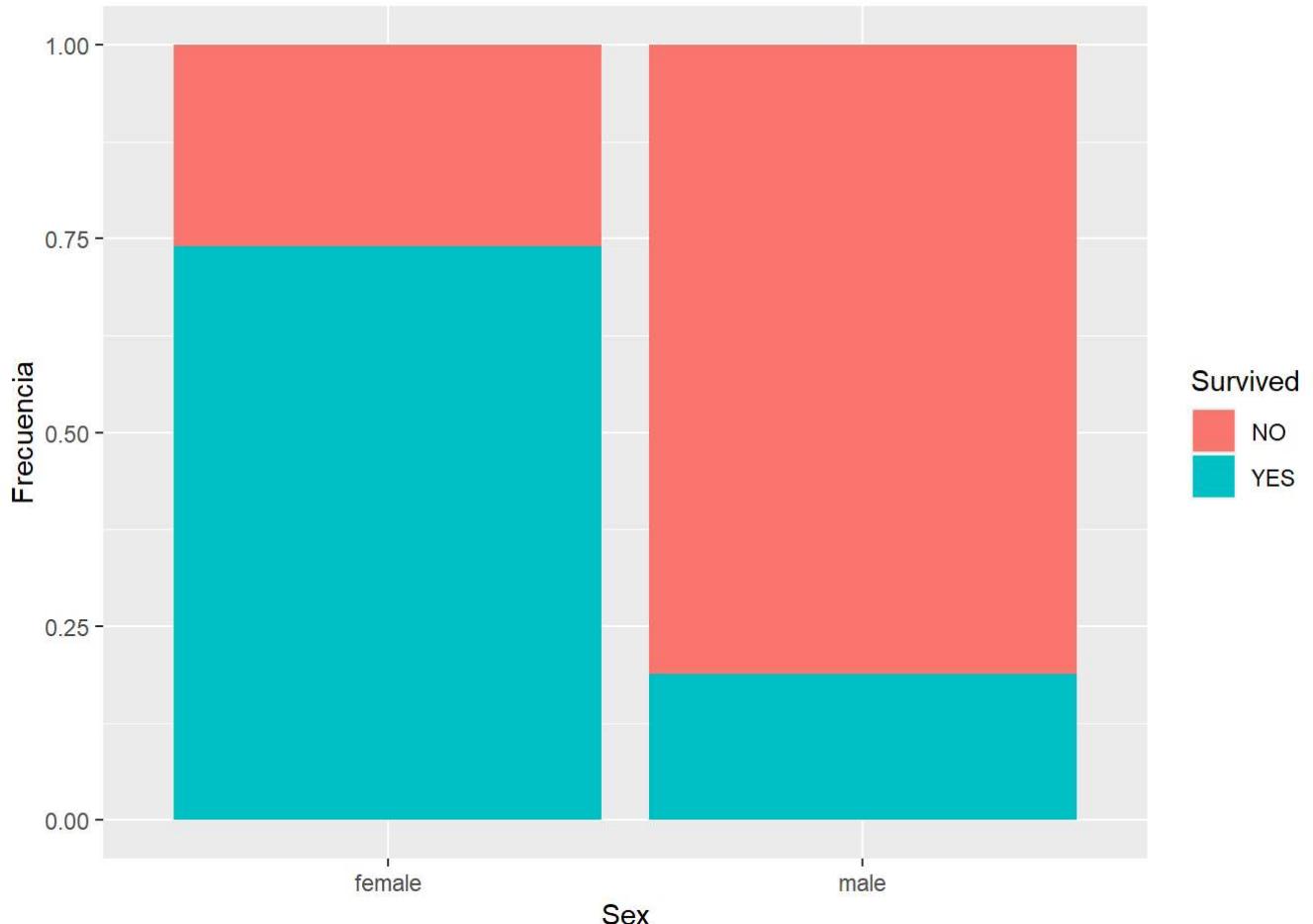


```
ggplot(data_selected,aes(x=Age,fill=Survived))+geom_histogram(binwidth =3)
```

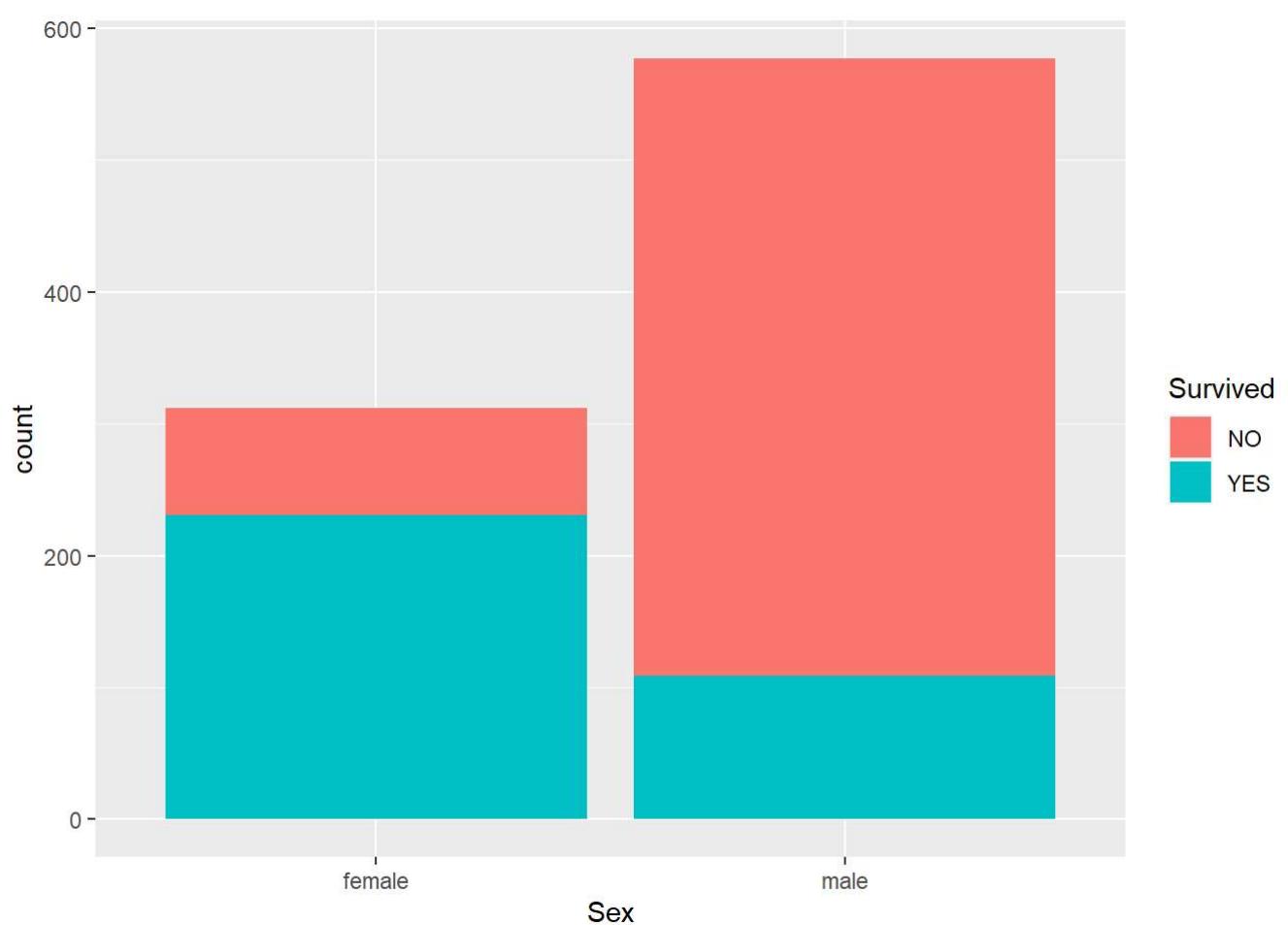


- Análisis por sexo: Podemos comprobar que en proporción sobrevivieron muchas más mujeres (casi un 75%) que hombres (menos del 20%). Por lo que se corrobora que se salvaron más mujeres que hombres.

```
ggplot(data = data_selected,aes(x=Sex,fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```

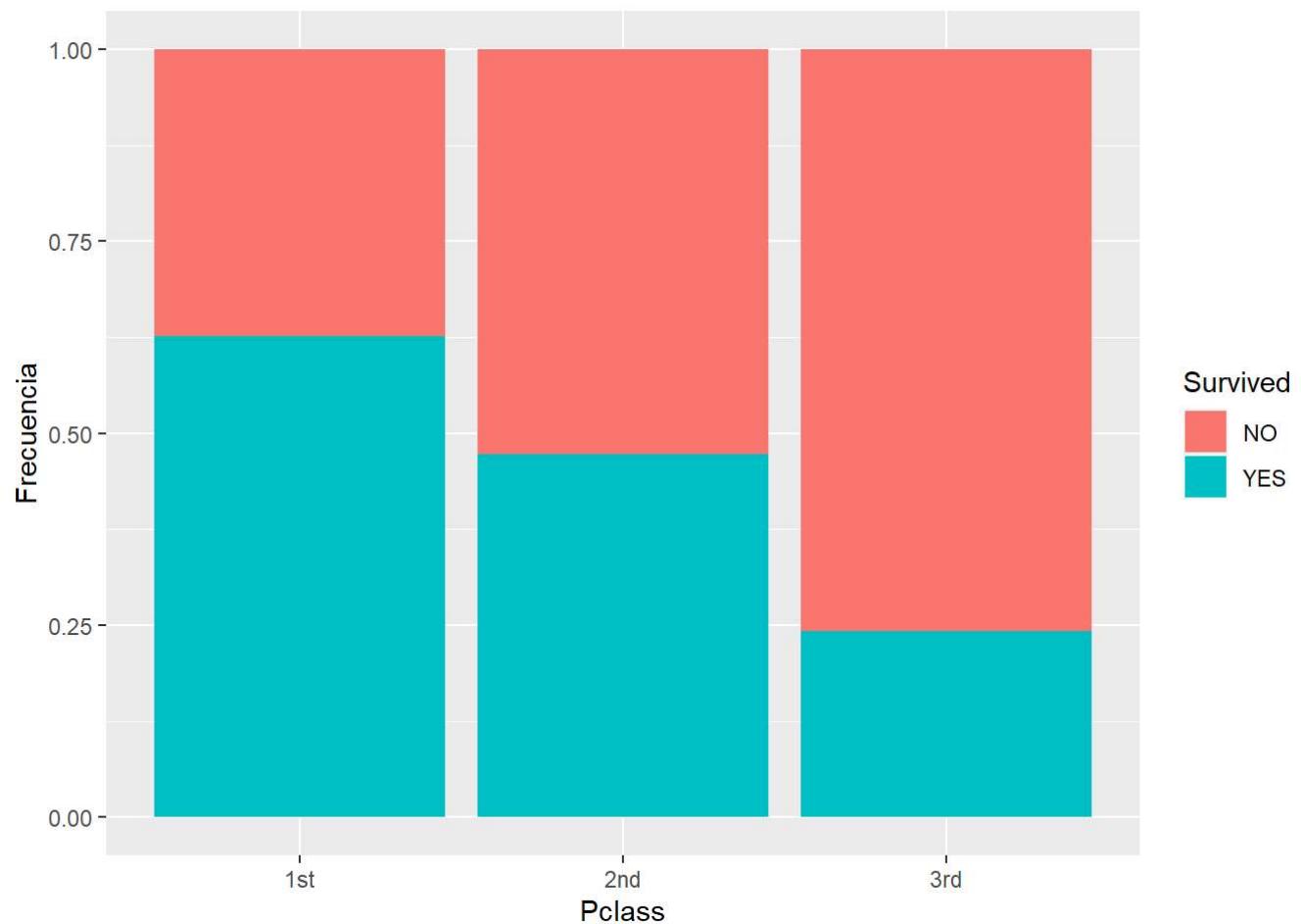


```
ggplot(data=data_selected,aes(x=Sex,fill=Survived))+geom_bar()
```

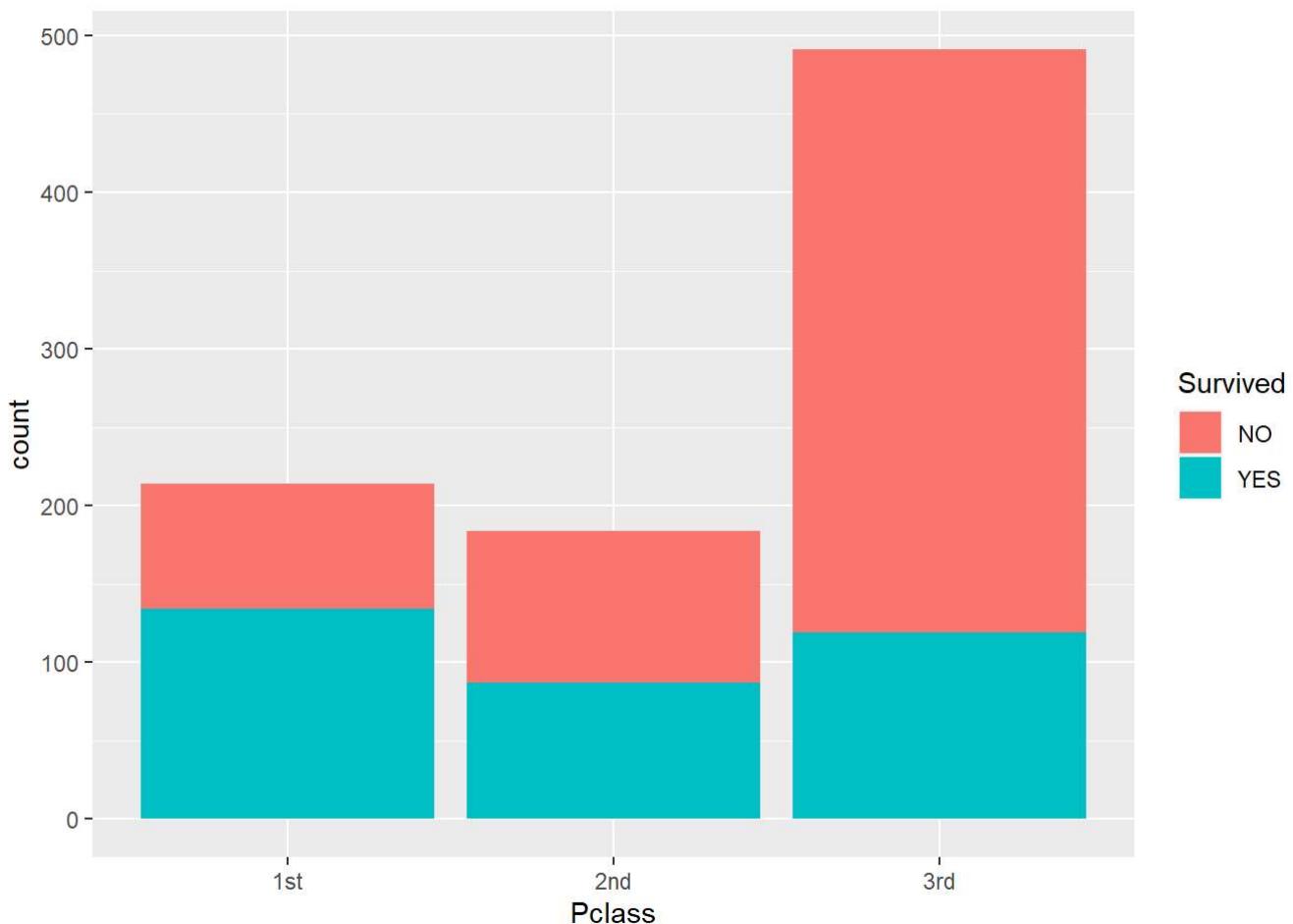


- Análisis por clase: Como se puede comprobar, menos del 25% de los que viajaban en tercera clase sobrevivieron.

```
ggplot(data = data_selected,aes(x=Pclass,fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```



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



- Análisis por edad, sexo y clase: vamos a hacer una análisis para comprobar la siguiente conclusión que hemos visto en el árbol de decisión:

- Si es mujer
  - no viaja en tercera clase, entonces sobrevive con una probabilidad del 93,86%
  - viaja en tercera clase
    - su edad es mayor o igual que 38,5 años, entonces muere (100%)
    - su edad es menor de 21,5 años, entonces no sobrevive al 58,82%
    - su edad es menor de 38,5 años pero igual o mayor a 21,5 años, entonces sobrevive al 62%

Cara a hacer esto, filtramos el dataset por **Sex** igual a "female" y la dividimos en 3 gráficas, una por clase utilizando la función **facet\_wrap**.

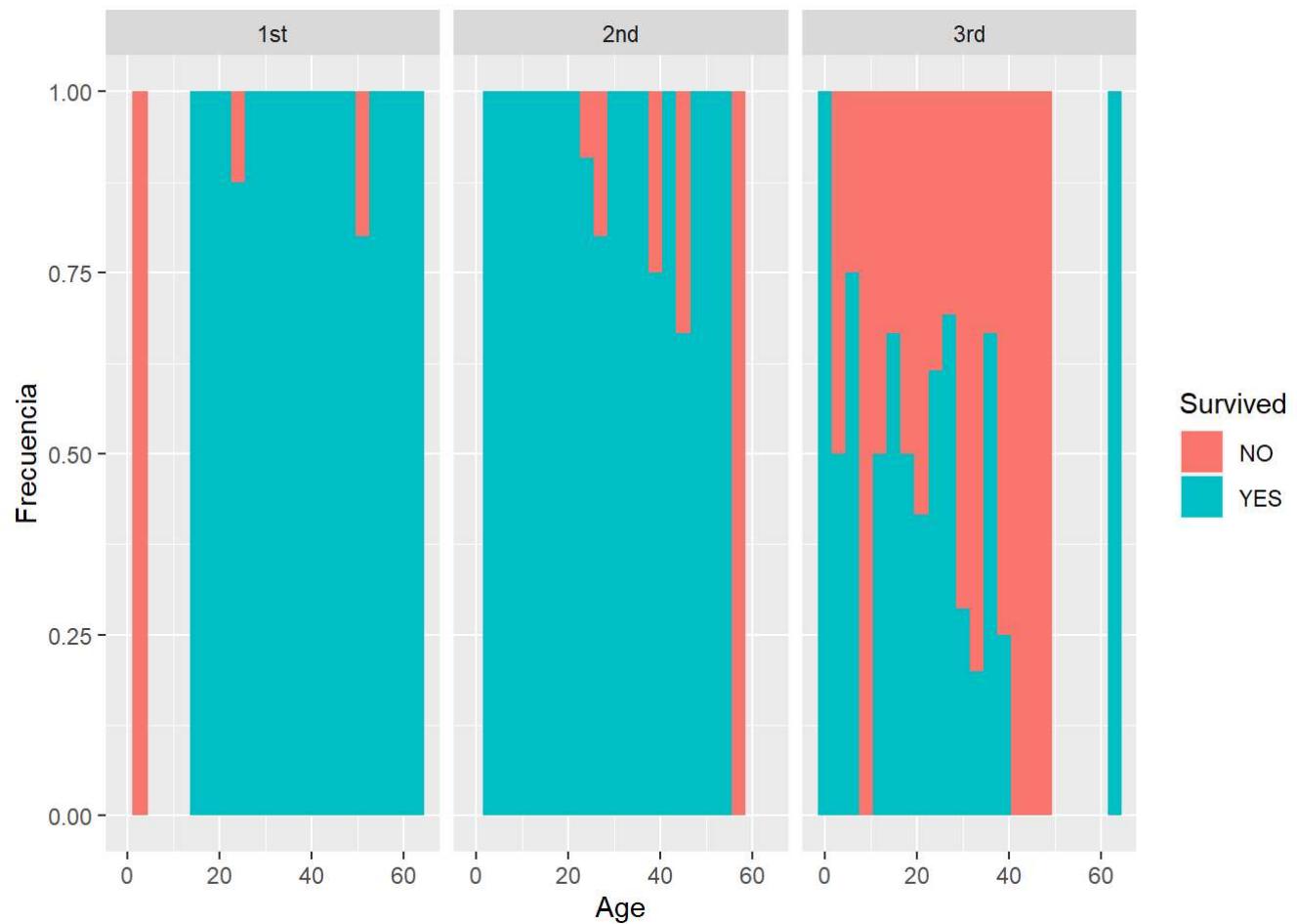
Tal y como se puede ver, la supervivencia en tercera clase para mujeres era nula a partir de los 40 años, con excepción de algún caso puntual.

Se puede ver que en el resto de franjas de edad el nivel de supervivencia podía estar en torno al 60%.

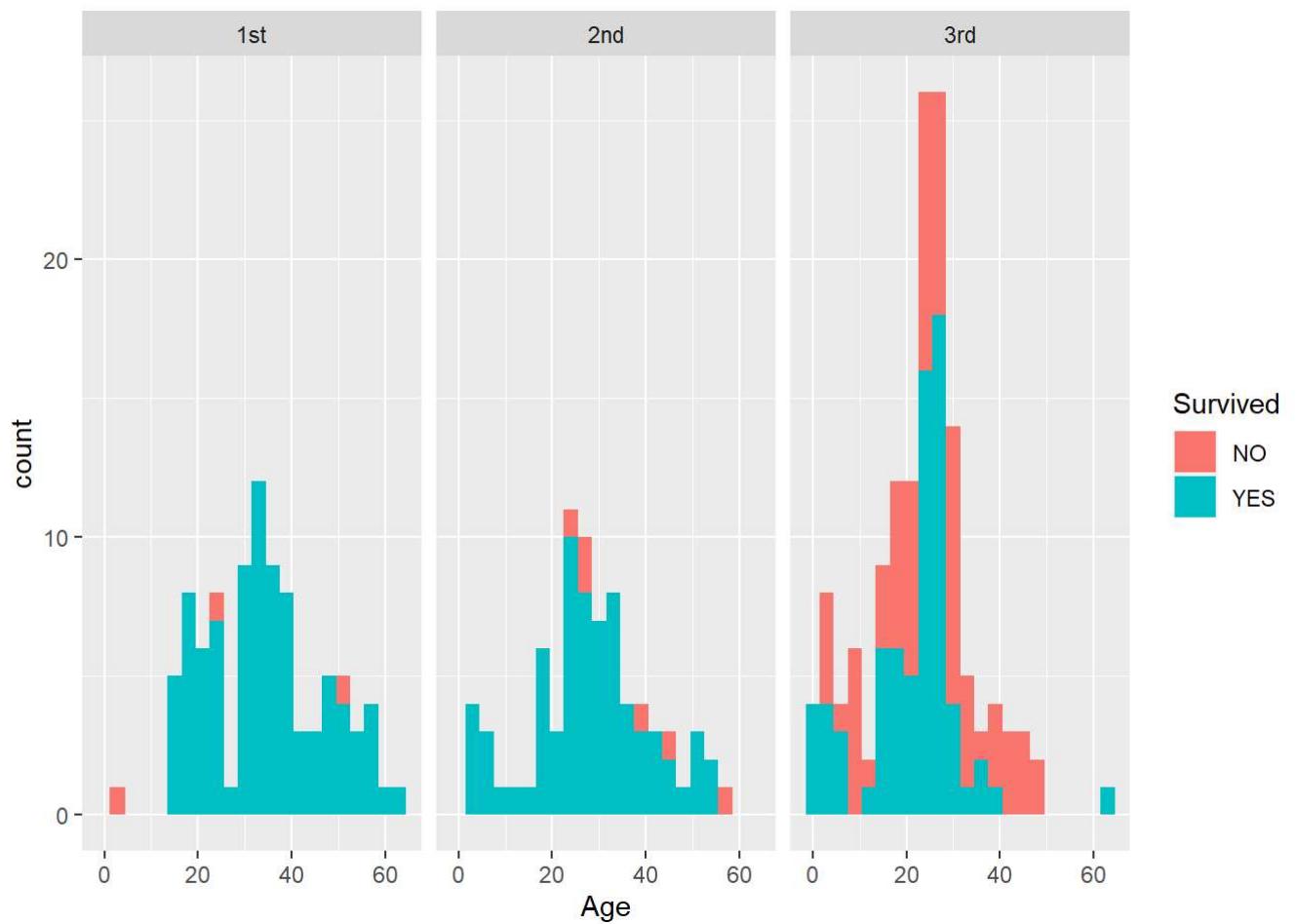
Un dato interesante es que se puede ver que incluso en tercera clase la mayoría de las niñas más pequeñas sobrevivieron, aunque sobre los 10 años la supervivencia fue nula.

También llama la atención que en primera clase las pocas niñas pequeñas que podrían haber, murieron todas.

```
ggplot(filter(data_selected, Sex=="female"),aes(x=Age,fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")+facet_wrap(~Pclass)+geom_histogram(binwidth =3, position = "fill")
```



```
ggplot(filter(data_selected, Sex=="female"),aes(x=Age,fill=Survived))+geom_bar()+facet_wrap(~Pclass)+geom_histogram(binwidth =3)
```



# 6. Conclusiones

---

A modo de resumen, hemos visto a lo largo del análisis los siguientes puntos:

- Hay variables como **Name** y **PassengerId** que no aportan información a priori al análisis y que hemos decidido eliminar.
- En el caso del tratamiento de valores vacíos, en algunos casos hemos podido eliminar directamente las observaciones al ser pocos casos (**embarked**), en otros hemos podido aplicar la media (**Age**) y otros casos se referían a pasajeros que no viajaban en cabina.
- No hemos encontrado outliers fuera de lo normal, estaban dentro del dominio.
- Hemos visto que **catplant**, la categoría por planta, no afectaba en cuanto a nivel de supervivencia.
- Hemos demostrado mediante contraste de hipótesis que de media los supervivientes son más jóvenes que quienes acabaron muriendo.
- Hemos visto también que hay una relación entre sexo y supervivencia por contraste de hipótesis.
- A través de una regresión lineal hemos visto cierta dependencia entre edad y supervivencia aunque era una dependencia débil.
- Mediante un árbol de decisión, hemos comprobado que se cumple la hipótesis inicial de que había más probabilidad de salvarse si se era mujer o niño y que viajar en tercera clase suponía una mayor probabilidad de morir que en el resto de clases.
- Con las correlaciones hemos visto que el precio del ticket tiene un nivel de relación muy bajo con respecto a la edad.
- Podemos ver por tanto, que la conocida cita de “Mujeres y niños primero”, se cumplía en general, seguida de un fuerte sesgo por clases económicas.

## 7. Contribuciones

<b>Contribuciones</b>	<b>Firma</b>
Investigación previa	Marco Emilio Rodríguez Serrano, Luis García Tarraga
Redacción de las respuestas	Marco Emilio Rodríguez Serrano, Luis García Tarraga
Desarrollo código	Marco Emilio Rodríguez Serrano, Luis García Tarraga