Práctica 2 - Limpieza y análisis de datos

Maria Dolores Moyano Guerrero y Victor Cancer Castillo

25 de Mayo de 2022

Contents


```
library(ggplot2)
library(corrplot)
library(faraway)
library(ggfortify)
library(ResourceSelection)
library(pROC)
library(grid)
library(colorspace)
library(Rcpp)
library(vctrs)
library(tidyverse)
library(VIM)
library(ggpubr)
library(caTools)
```

Titanic: Machine Learning from Disaster

Descripción del dataset

El desastre del RMS Titanic fue un accidente marítimo que acaeció en el 1912 y que se llevó por delante más de 1500 vidas. A bordo del Titanic iban más de 2000 pasajeros, por lo que cerca del 75% de los pasajerons fallecieron en el hundimiento del barco el cual no tenia botes salvavidas para todos los pasajeros.

Estas muertes no se dieron por igual para todos los grupos de pasajeros de manera aleatoria, sino que parece ser que hubo grupos dentro del barco que tuvieron más probabilidad de morir que otros, como podremos ver en este estudio.

Nos vamos a centrar aquí en tratar de averiguar qué características compartían en común los pasajeros que se salvaron/fallecieron para tratar de crear un modelo que sea capaz de predecir si un pasajero iba a morir o no.

Integración y selección de los datos

Para tratar este problema vamos a utilizar los datos que se ofrecen en la competicción de Kaggle, donde se da un dataset que contiene datos para entrenar el modelo y otro para hacer los tests del modelo creado.

Por un lado tenemos los datos para entrenar el modelo

```
train <- read.table(file="train.csv",sep=',',dec='.',stringsAsFactors = TRUE,header=TRUE)
summary(train)
```

```
##   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
```

Y por otro tenemos los datos para testear dicho modelo

```
test <- read.table(file="test.csv",sep=',',dec='.',stringsAsFactors = TRUE,header=TRUE)
summary(test)
```

```
##   PassengerId         Pclass
##  Min.   : 892.0   Min.   :1.000
##  1st Qu.: 996.2   1st Qu.:1.000
##  Median :1100.5   Median :3.000
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##                                          Name         Sex           Age
##  Abbott, Master. Eugene Joseph             :  1   female:152   Min.   : 0.17
##  Abelseth, Miss. Karen Marie               :  1   male  :266   1st Qu.:21.00
##  Abelseth, Mr. Olaus Jorgensen             :  1                Median :27.00
##  Abrahamsson, Mr. Abraham August Johannes  :  1                Mean   :30.27
##  Abrahim, Mrs. Joseph (Sophie Halaut Easu):  1                3rd Qu.:39.00
##  Aks, Master. Philip Frank                 :  1                Max.   :76.00
##  (Other)                                   :412                NA's   :86
##      SibSp            Parch             Ticket         Fare
##  Min.   :0.0000   Min.   :0.0000   PC 17608:  5   Min.   :  0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   113503  :  4   1st Qu.:  7.896
##  Median :0.0000   Median :0.0000   CA. 2343:  4   Median : 14.454
##  Mean   :0.4474   Mean   :0.3923   16966   :  3   Mean   : 35.627
##  3rd Qu.:1.0000   3rd Qu.:0.0000   220845  :  3   3rd Qu.: 31.500
##  Max.   :8.0000   Max.   :9.0000   347077  :  3   Max.   :512.329
##                                    (Other) :396   NA's   :1
##              Cabin     Embarked
##                 :327   C:102
##  B57 B59 B63 B66:  3   Q: 46
##  A34            :  2   S:270
##  B45            :  2
##  C101           :  2
##  C116           :  2
##  (Other)        : 80
```

Las variables que incluye el dataset son las siguientes:

- *PassengerId*: Número de identificación del pasajero
- *Survived*: Indica si el pasajero sobrevivió (0 = No, 1 = Sí)
- *Pclass*: Clase de tiquet (1 = Primera clase, 2 = Segunda clase, 3 = Tercera clase)
- *Name*: Nombre del pasajero
- *Sex*: Sexo del pasajero
- *Age*: Edad del pasajero
- *SibSp*: Número de hermanos/hermanas, esposos/esposas a bordo del Titanic
- *Parch*: Número de padres/madres, hijos/hijas a bordo del Titanic
- *Ticket*: Número de ticket

- *Fare*: Tarifa del pasajero
- *Cabin*: Número de cabina
- *Embarked*: Puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton)

Para hacer análisis (no modelaje) trataremos los datos completos (es decir los datos de test y de entrenamiento, sin la columna *Survived*)

```
full <- rbind(test,train[-which(names(train) == "Survived")])
```

Limpieza de los datos

En primer lugar, vamos a estudiar si los datos tienen elementos vacíos

Elementos nulos o ceros

Embarked

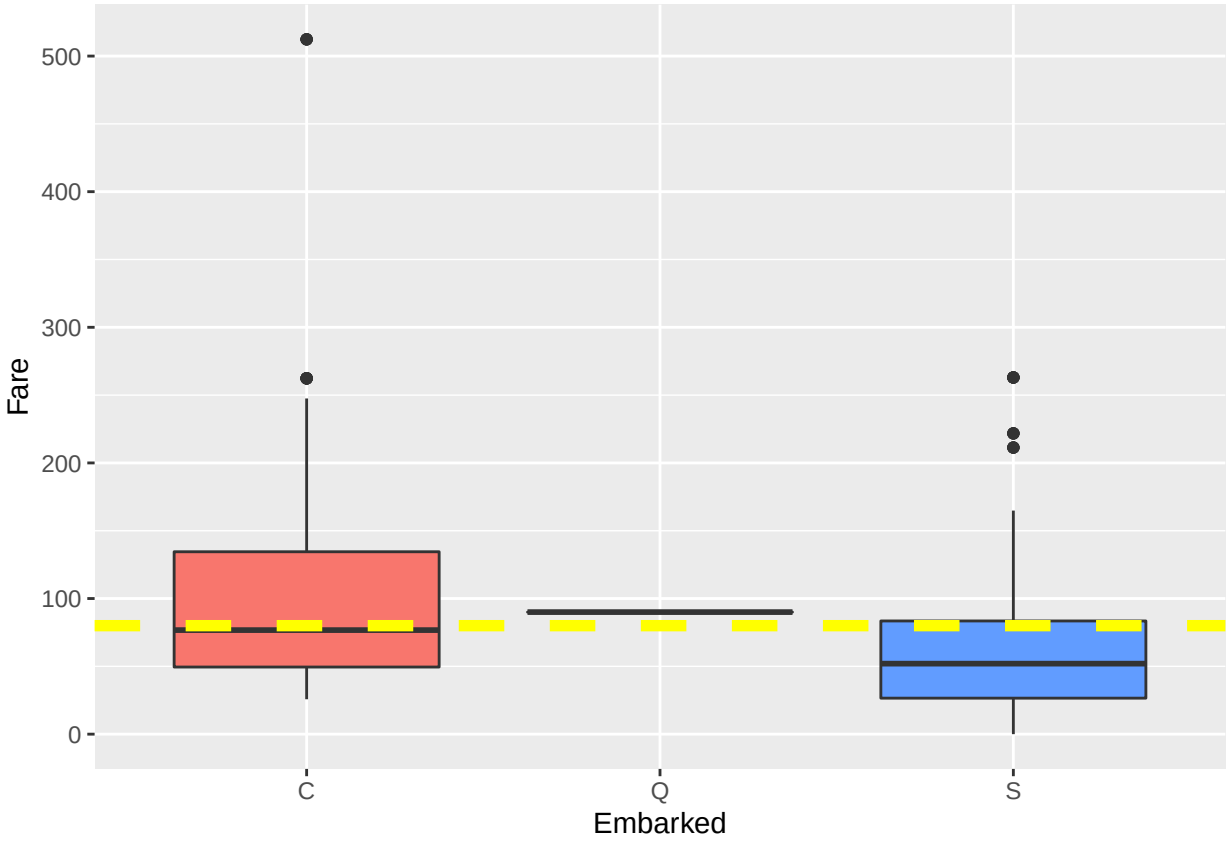
Vemos entre los valores de la columna Embarked del dataset de entrenamiento que hay dos valores vacíos

```
full[full$Embarked == "",]
```

```
##      PassengerId Pclass                                       Name    Sex Age
## 480           62      1                        Icard, Miss. Amelie female  38
## 1248         830      1 Stone, Mrs. George Nelson (Martha Evelyn) female  62
##      SibSp Parch Ticket Fare Cabin Embarked
## 480      0     0 113572   80   B28
## 1248     0     0 113572   80   B28
```

Probablemente la relación más relevante entre el puerto de embarque la tiene el precio del billete (pues al hacer un viaje más largo se cobrará más al pasajero). Por lo tanto veamos con qué puerto encajan más estas dos pasajeras sabiendo que ellas pagaron 80$ por su billete de primera clase:

```
ggplot(full[full$Embarked != "" & full$Pclass == "1",],aes(x=Embarked,y=Fare, fill=Embarked)) + geom_bo
    theme(legend.position="none") + geom_hline(aes(yintercept=80), colour='yellow', linetype='dashed',
```



De esta gráfica podemos deducir que estas mujeres probablemente embarcaron en el puerto C, así que imputaremos ese valor a ambas mujeres:

```
full[full$Embarked=="",]$Embarked <- "C"
train[train$Embarked=="",]$Embarked <- "C"
```

Fare

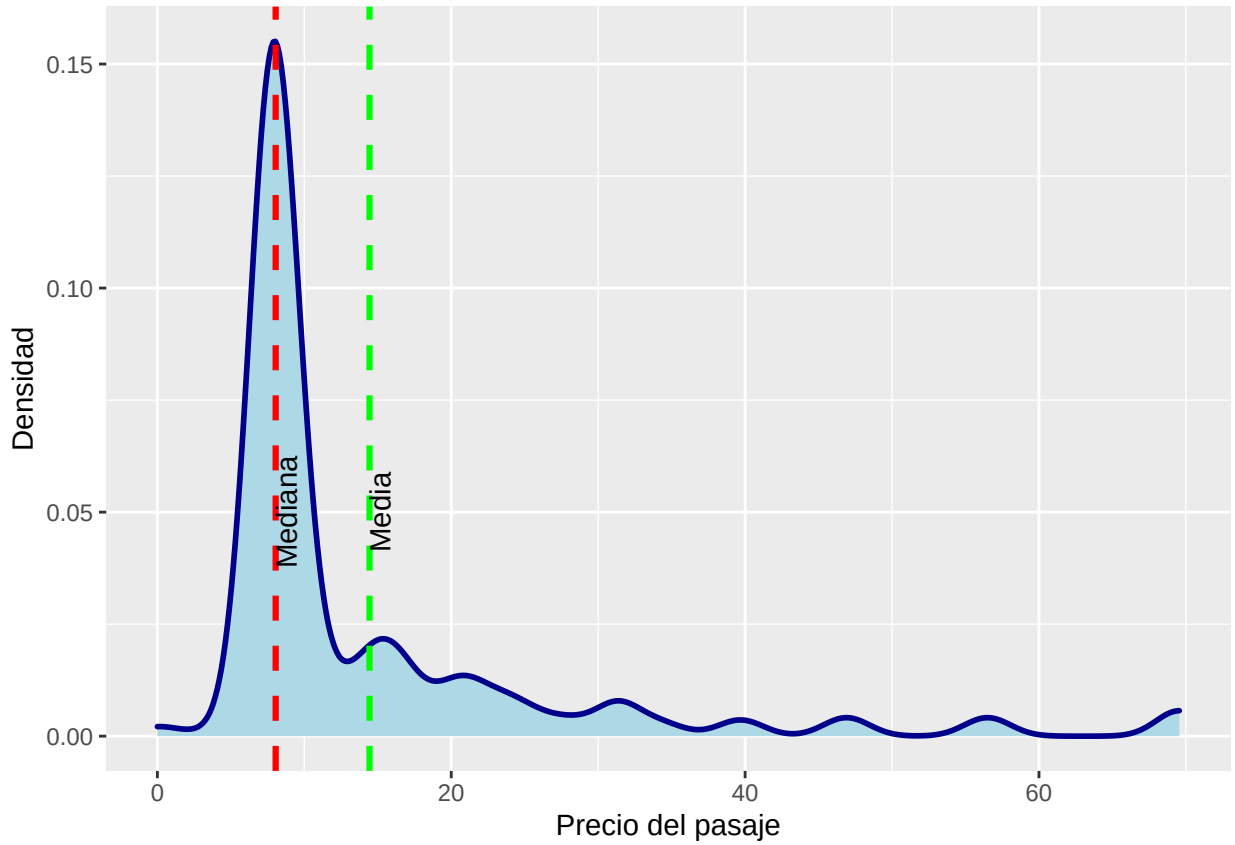
De las tarifas de los pasajes encontramos que tan solo hay un caso donde desconocemos el precio que se pagó:

```
full[is.na(full$Fare),]
```

```
##     PassengerId Pclass               Name  Sex  Age SibSp Parch Ticket Fare
## 153        1044      3 Storey, Mr. Thomas male 60.5     0     0   3701   NA
##     Cabin Embarked
## 153              S
```

De nuevo vamos a observar cuanto costaron estos pasajes observando el puerto de embarcación y la clase a la que pertenece este pasajero

```
ggplot(full[!is.na(full$Fare) & full$Pclass == "3" & full$Embarked == "S" ,], aes(x=Fare)) +
  geom_density(color="darkblue", fill="lightblue",size=1)+ylab("Densidad")+xlab("Precio del pasaje") +
  geom_vline(xintercept = median(full[!is.na(full$Fare) & full$Pclass == "3" & full$Embarked == "S" ,]$F
  geom_vline(xintercept = mean(full[!is.na(full$Fare) & full$Pclass == "3" & full$Embarked == "S" ,]$Far
  annotate(geom = "text", label = c("Mediana", "Media"), x = c(median(full[!is.na(full$Fare) & full$Pcla
```



Viendo la distribución de los datos vemos que lo más correcto sería coger la mediana del precio del pasaje, que en este caso es 8.05

```
fare_median <-  median(full[!is.na(full$Fare) & full$Pclass == "3" & full$Embarked == "S" ,]$Fare)

full[is.na(full$Fare),]$Fare <- fare_median
test[is.na(test$Fare),]$Fare <- fare_median
```

Por otro lado tenemos registros donde el precio del pasaje fue cero

```
full[full$Fare == 0,]
```

```
##      PassengerId Pclass                                 Name  Sex Age SibSp
## 267         1158      1 Chisholm, Mr. Roderick Robert Crispin male  NA     0
## 373         1264      1               Ismay, Mr. Joseph Bruce male  49     0
## 598          180      3                   Leonard, Mr. Lionel male  36     0
## 682          264      1                Harrison, Mr. William male  40     0
## 690          272      3         Tornquist, Mr. William Henry male  25     0
## 696          278      2          Parkes, Mr. Francis "Frank" male  NA     0
## 721          303      3       Johnson, Mr. William Cahoone Jr male  19     0
## 832          414      2        Cunningham, Mr. Alfred Fleming male  NA     0
## 885          467      2                Campbell, Mr. William male  NA     0
## 900          482      2      Frost, Mr. Anthony Wood "Archie" male  NA     0
## 1016         598      3                   Johnson, Mr. Alfred male  49     0
## 1052         634      1         Parr, Mr. William Henry Marsh male  NA     0
```

```
## 1093         675      2            Watson, Mr. Ennis Hastings male  NA     0
## 1151         733      2                 Knight, Mr. Robert J male  NA     0
## 1225         807      1               Andrews, Mr. Thomas Jr male  39     0
## 1234         816      1                     Fry, Mr. Richard male  NA     0
## 1241         823      1       Reuchlin, Jonkheer. John George male  38     0
##      Parch Ticket Fare       Cabin Embarked
## 267      0 112051    0                   S
## 373      0 112058    0 B52 B54 B56       S
## 598      0   LINE    0                   S
## 682      0 112059    0         B94       S
## 690      0   LINE    0                   S
## 696      0 239853    0                   S
## 721      0   LINE    0                   S
## 832      0 239853    0                   S
## 885      0 239853    0                   S
## 900      0 239854    0                   S
## 1016     0   LINE    0                   S
## 1052     0 112052    0                   S
## 1093     0 239856    0                   S
## 1151     0 239855    0                   S
## 1225     0 112050    0         A36       S
## 1234     0 112058    0        B102       S
## 1241     0  19972    0                   S
```

Haciendo una busqueda por internet de los nombres de algunas de estas personas vemos algo que podíamos sopechar: eran parte de los trabajadores de la embarcación o relacionados con ésta (como el propio diseñador del Titanic, Roderick Robert Crispin).

Puesto que realmente el pasaje no valía cero dolares sino que estas personas fueron invitadas, lo que vamos a hacer para que ésto no desvirtue los datos es imputar de nuevo la median, en este caso lo haremos según la clase de pasaje que tuvieran (todos eran del puerto de embarcación S)

```
median_fare_1 <- median(full[full$Fare != 0 & full$Pclass == 1 & full$Embarked == 'S',]$Fare)
median_fare_2 <- median(full[full$Fare != 0 & full$Pclass == 2 & full$Embarked == 'S',]$Fare)
median_fare_3 <- median(full[full$Fare != 0 & full$Pclass == 3 & full$Embarked == 'S',]$Fare)

#Imputamos según la clase en los dataset que hemos generado:
full[full$Fare == 0 & full$Pclass == 1,]$Fare <- median_fare_1
full[full$Fare == 0 & full$Pclass == 2,]$Fare <- median_fare_2
full[full$Fare == 0 & full$Pclass == 3,]$Fare <- median_fare_3

train[train$Fare == 0 & train$Pclass == 1,]$Fare <- median_fare_1
train[train$Fare == 0 & train$Pclass == 2,]$Fare <- median_fare_2
train[train$Fare == 0 & train$Pclass == 3,]$Fare <- median_fare_3

test[test$Fare == 0 & test$Pclass == 1,]$Fare <- median_fare_1

#Los siguientes casos no existen en el dataset de test:
#test[test$Fare == 0 & test$Pclass == 2,]$Fare <- median_fare_2
#test[test$Fare == 0 & test$Pclass == 3,]$Fare <- median_fare_3
```

Age

En la variable de edad encontramos que hay 177 NAs en el dataset de entrenamiento y 86 NAs en el de test.

La edad es una variable algo más complicada de imputar y una opción sería utilizar la mediana de la edad de los pasajeros, pero vamos a optar por utilizar el metodo kNN que nos imputará el valor de la edad utilizando los valores de los puntos más cercanos al que nos falta.

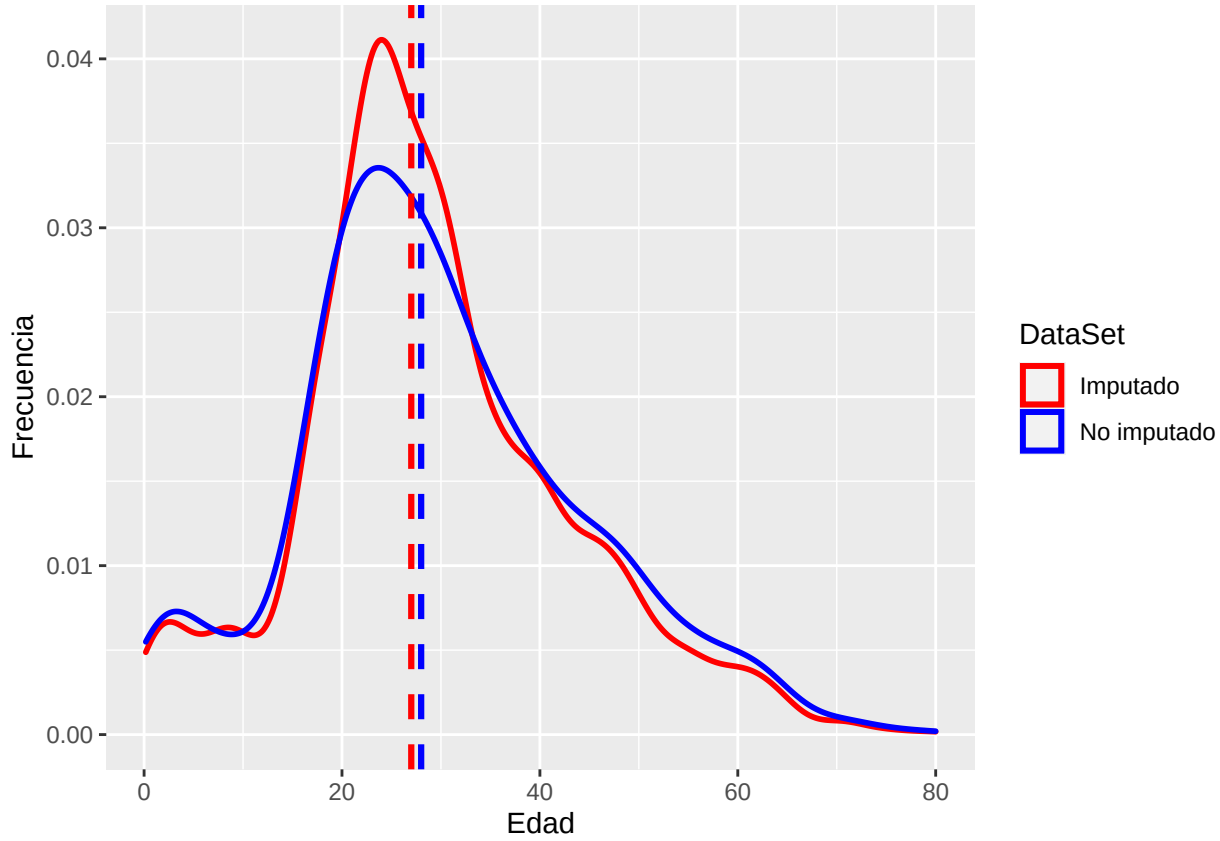
Las variables que tendremos en cuenta en esta imputación serán:

- Sex
- PClass
- SibSp
- Parch
- Fare
- Embarked

```
full_imp <- kNN(full,k=11,dist_var=c('Sex','Pclass','SibSp','Fare','Parch','Embarked'),variable='Age')
```

Para ver si esta imputación ha afectado a la distribución de edad

```
ggplot() +
  geom_density(data=full_imp, aes(x=Age,color='Imputado') , size=1) +
  geom_density(data=full, aes(x=Age, color = 'No imputado') ,size=1) +
  geom_vline(xintercept = median(full$Age,na.rm = TRUE),color="blue",size=1.1,linetype="dashed") +
  geom_vline(xintercept = median(full_imp$Age),color="red",size=1.1,linetype="dashed") +
  ylab("Frecuencia") + xlab("Edad") + theme(legend.position = 'right') +
  scale_color_manual("DataSet",values = c('Imputado' = 'red', 'No imputado' = 'blue'))
```



Podemos ver un crecimiento en la densidad de valores alrededor de la mediana, pero la distribución sigue teniendo una forma parecida a la de ante de imputar valores, por lo que damos por correctos los datos que hemos introducido para los valores NA de la edad.

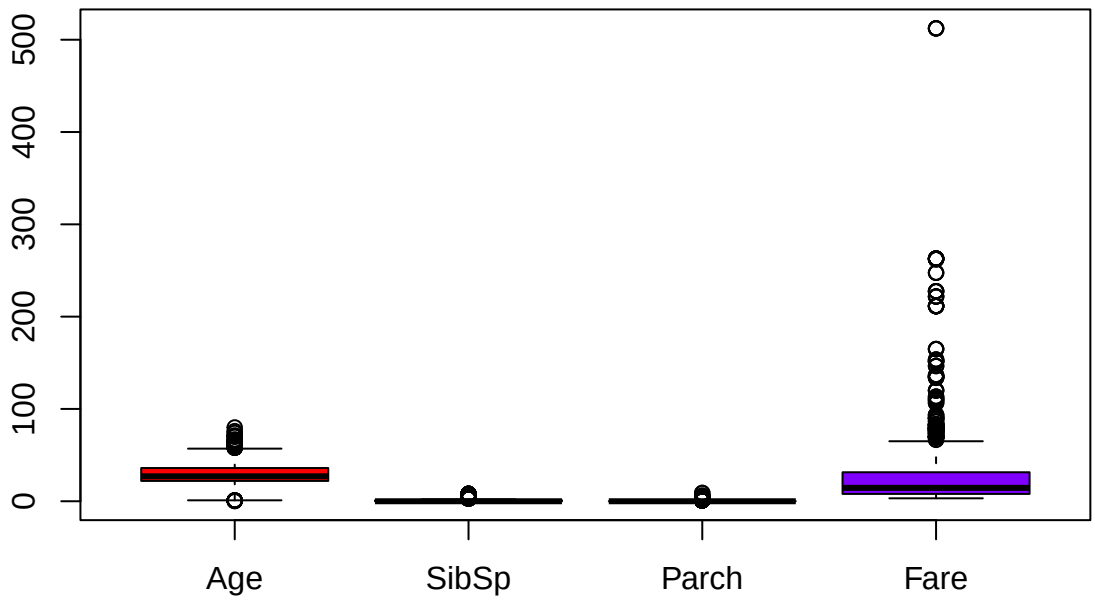
Por lo tanto pasamos ahora a imputar estos valores en los datasets que estamos ahora gestionando:

```
full$Age <- full_imp$Age

train <- merge(train, full_imp[c('PassengerId','Age')], by.x=c("PassengerId"), by.y=c("PassengerId"), al
train <- train[,-which(names(train) %in% c("Age.x","PassengerId.y"))]
train %>% rename( Age = Age.y )

test <- merge(test, full_imp[c('PassengerId','Age')], by.x=c("PassengerId"), by.y=c("PassengerId"), all
test <- test[,-which(names(test) %in% c("Age.x","PassengerId.y"))]
test %>% rename( Age = Age.y )
```

Outliers

Los valores extremos (o outliers) son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población. Hay diferentes métodos para identificar valores extremos, uno de ellos es mediante gráficos de cajas (boxplots), otros se basan en la distancia de Mahanlanobis o distancia de Cook, también se usan modelos estadísticos, supervisados o no supervisados, por ejemplo, mediante técnicas de clustering. En este caso utilizaremos la función boxplots.stats() de R.

```
borrar<-c("PassengerId","Name","Ticket","Pclass","Embarked","Survived","Sex","Cabin" )
fullr<-full[,!names(full) %in% borrar]
boxplot(fullr, col=rainbow(ncol(fullr)))
```



Los valores extremos detectados son:

Edad: Tras revisar los valores, se considera que son valores válidos

```
  boxplot.stats(full$Age)$out
```

```
##  [1] 62.00 63.00 60.00 60.00 67.00 76.00 63.00 61.00 60.50 64.00 61.00  0.33
## [13] 60.00 64.00  0.92  0.75 64.00  0.83 58.00  0.17 59.00 58.00 66.00 65.00
## [25]  0.83 59.00 71.00 70.50 61.00 58.00 59.00 62.00 58.00 63.00 65.00  0.92
## [37] 61.00 60.00 64.00 65.00  0.75 63.00 58.00 71.00 64.00 62.00 62.00 60.00
## [49] 61.00 80.00  0.75 58.00 70.00 60.00 60.00 70.00  0.67  0.42 62.00  0.83
## [61] 74.00
```

Fare (tarifa del pasajero):

```
boxplot.stats(full$Fare)$out
```

```
##   [1]  82.2667 262.3750  76.2917 263.0000 262.3750 262.3750 263.0000 211.5000
##   [9] 211.5000 221.7792  78.8500 221.7792  75.2417 151.5500 262.3750  83.1583
##  [17] 221.7792  83.1583  83.1583 247.5208  69.5500 134.5000 227.5250  73.5000
##  [25] 164.8667 211.5000  71.2833  75.2500 106.4250 134.5000 136.7792  75.2417
##  [33] 136.7792  82.2667  81.8583 151.5500  93.5000 135.6333 146.5208 211.3375
##  [41]  79.2000  69.5500 512.3292  73.5000  69.5500  69.5500 134.5000  81.8583
##  [49] 262.3750  93.5000  79.2000 164.8667 211.5000  90.0000 108.9000  71.2833
##  [57] 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000 263.0000
```

```
##  [65]  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500  69.5500
##  [73] 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000  79.2000
##  [81]  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500  91.0792
##  [89] 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750 164.8667
##  [97] 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000 263.0000
## [105]  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000 120.0000
## [113] 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792  90.0000
## [121]  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250  71.0000
## [129] 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000  78.2667
## [137] 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500  73.5000
## [145] 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250 211.3375
## [153] 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583 211.3375
## [161]  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500  89.1042
## [169] 164.8667  69.5500  83.1583
```

Se ha buscado el rango de precios de los billetes (https://www.20minutos.es/noticia/1365526/0/titanic/hundimiento/aniversario/), y los precios máximos de la gráfica, 512.32, están dentro del rango, con lo que se consideran valores válidos.

Análisis de los datos

En primer lugar, se va a dividir el conjunto de entrenamiento en varios grupos para realizar el análisis de los datos y así poder estudiar la supervivencia.

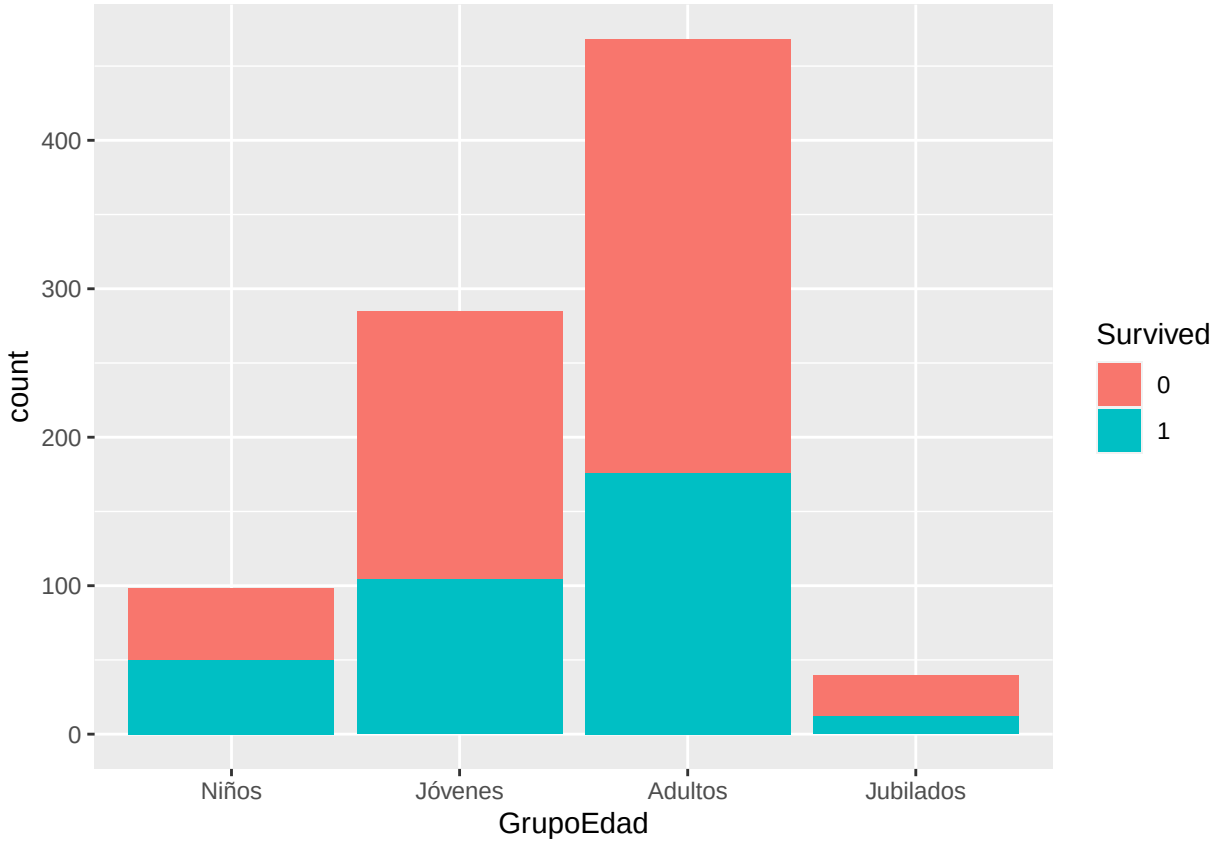
Selección de los grupos

Los grupos seleccionados serán los siguientes, para estudiar su relación con survived:

Age: se estudiará el efecto del rango de edad del pasajero en la supervivencia. Embarked: se analizará el efecto del puerto de embarque en la supervivencia. Parch: número de padres/madres, hijos/hijas a bordo del Titanic y su influencia. Pclass: se analizará la influencia de clase del pasajero. Sex: incluencia del sexo del pasajero en la supervivencia. SibSp: influencia del número de hermanos/hermanas, esposos/esposas a bordo del Titanic en la supervivencia.

Edad vs survived

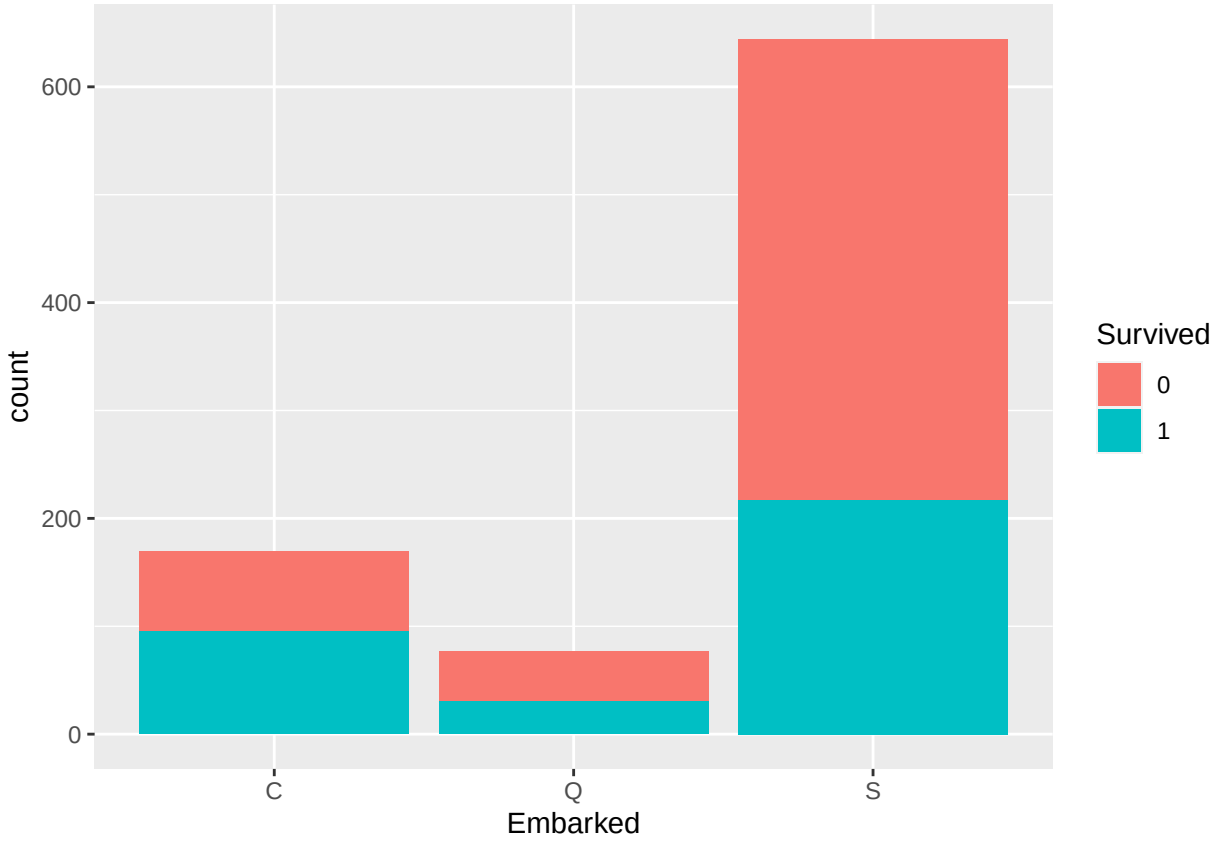
```
train$GrupoEdad <- cut(train$Age, breaks = c(0,15,25,55,100), labels = c("Niños","Jóvenes","Adultos","J
train$Survived <- as.factor(train$Survived)
PGedad<-ggplot(train, aes(x=GrupoEdad, fill=Survived)) + geom_bar()
PGedad
```



Se aprecia que el porcentaje de supervivientes aumenta cuanto menor es la edad.

Puerto de embarque vs survived

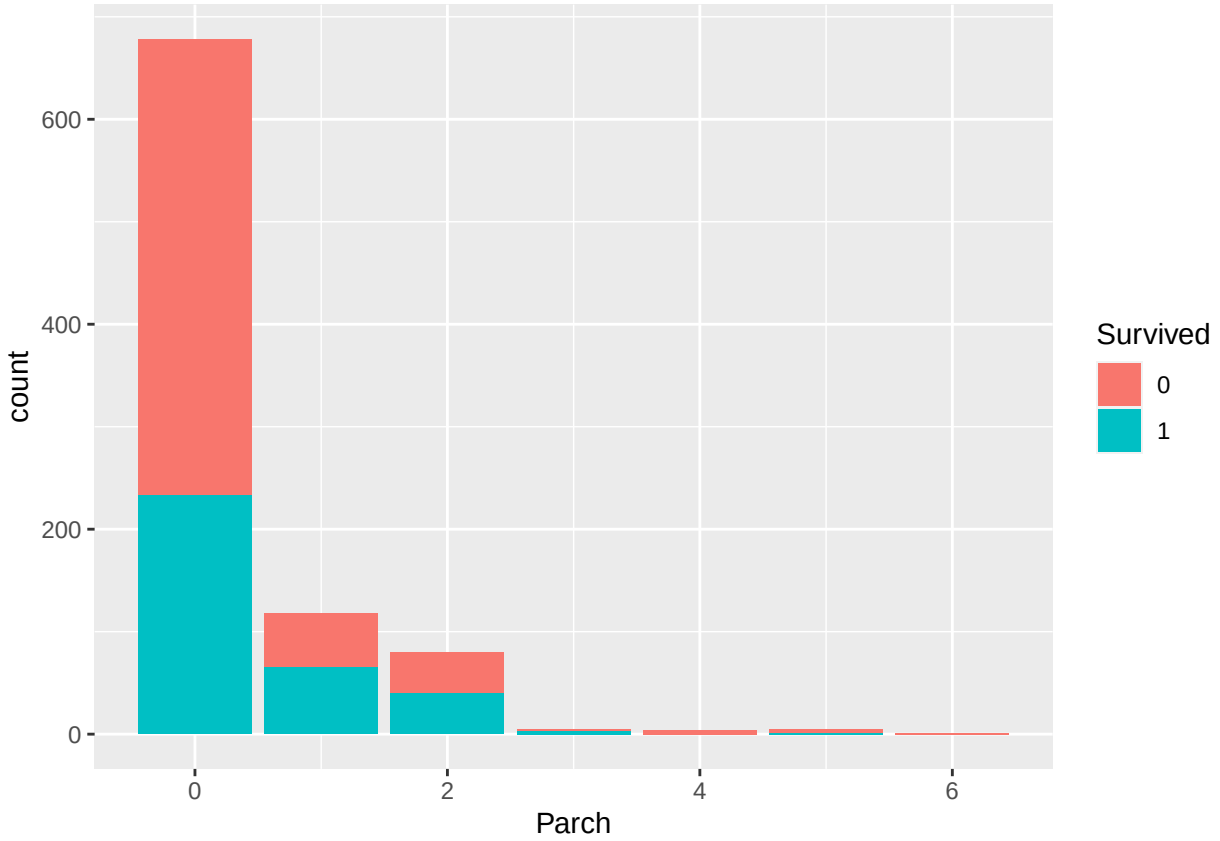
```
ggplot(train, aes(x=Embarked, fill=Survived)) + geom_bar()
```



Hay una menor tasa de supervivencia, de los pasajeros embarcados en Southampton con respecto a los embarcados en Cherbourg y Queenstown.

Padres/madres, hijos e hijas vs survived

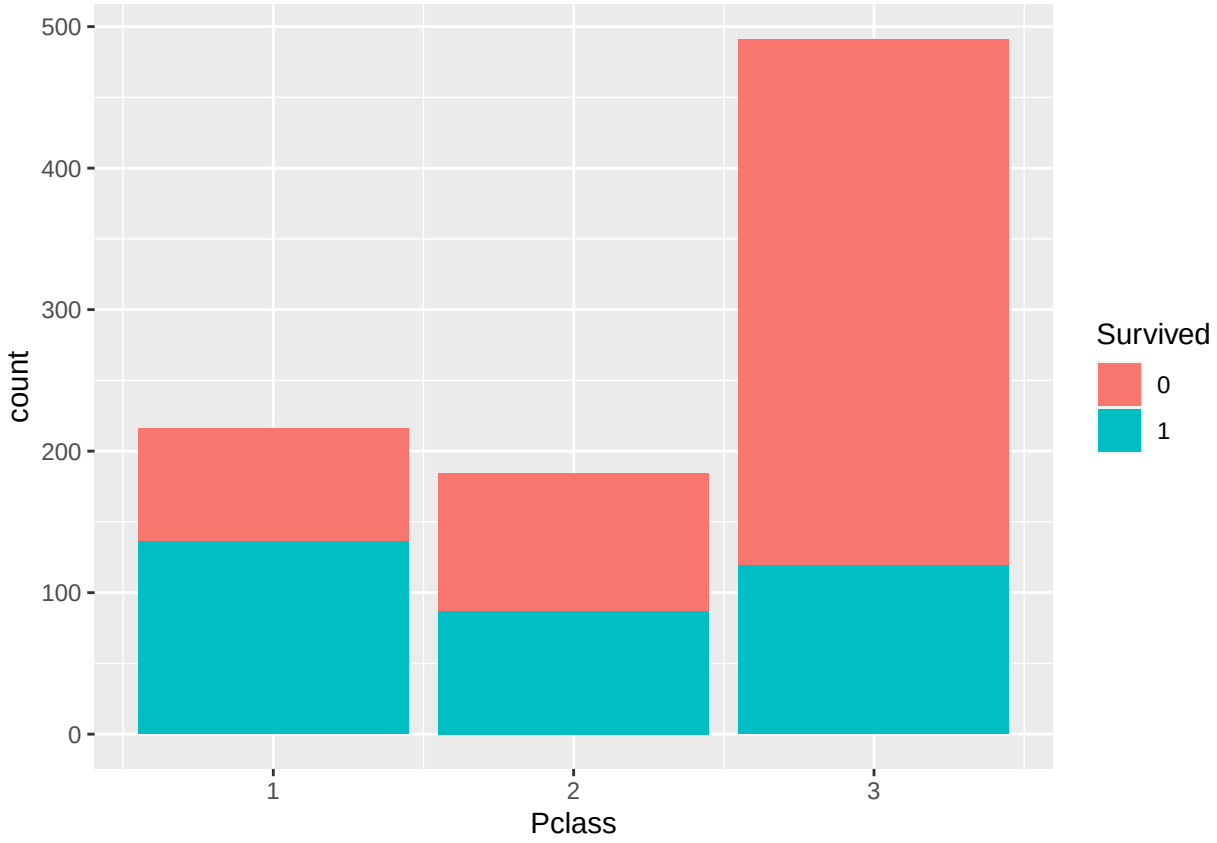
```
ggplot(train, aes(x=Parch, fill=Survived)) + geom_bar()
```



No se aprecia diferencias significativas entre los diferentes elementos de esta agrupación, entendiéndosa que el número de padres/madres, hijos e hijas no afecta a la supervivencia.

Clase del pasaje vs survived

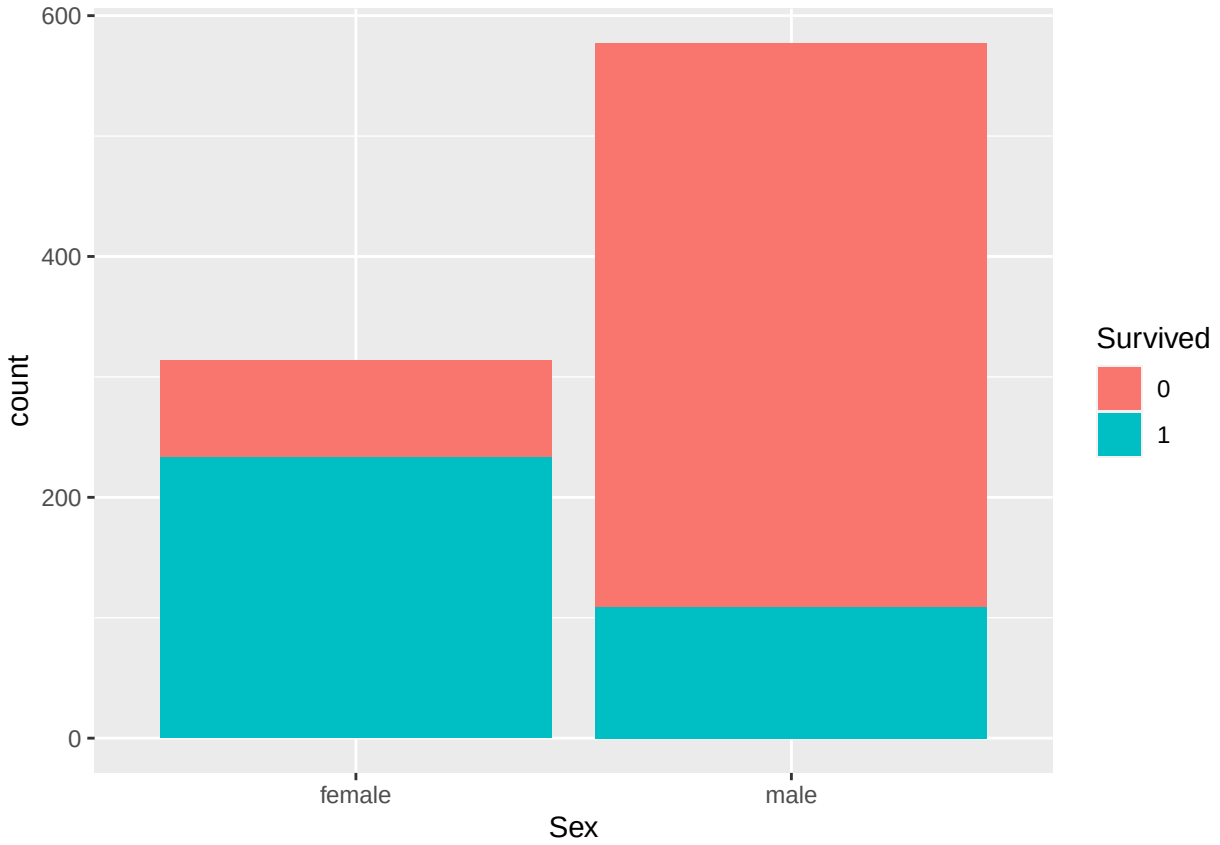
```
PClase<-ggplot(train, aes(x=Pclass, fill=Survived)) + geom_bar()
PClase
```



La clase es una variable que impacta fuertemente sobre la tasa de supervivencia.

Sexo del pasaje vs survived

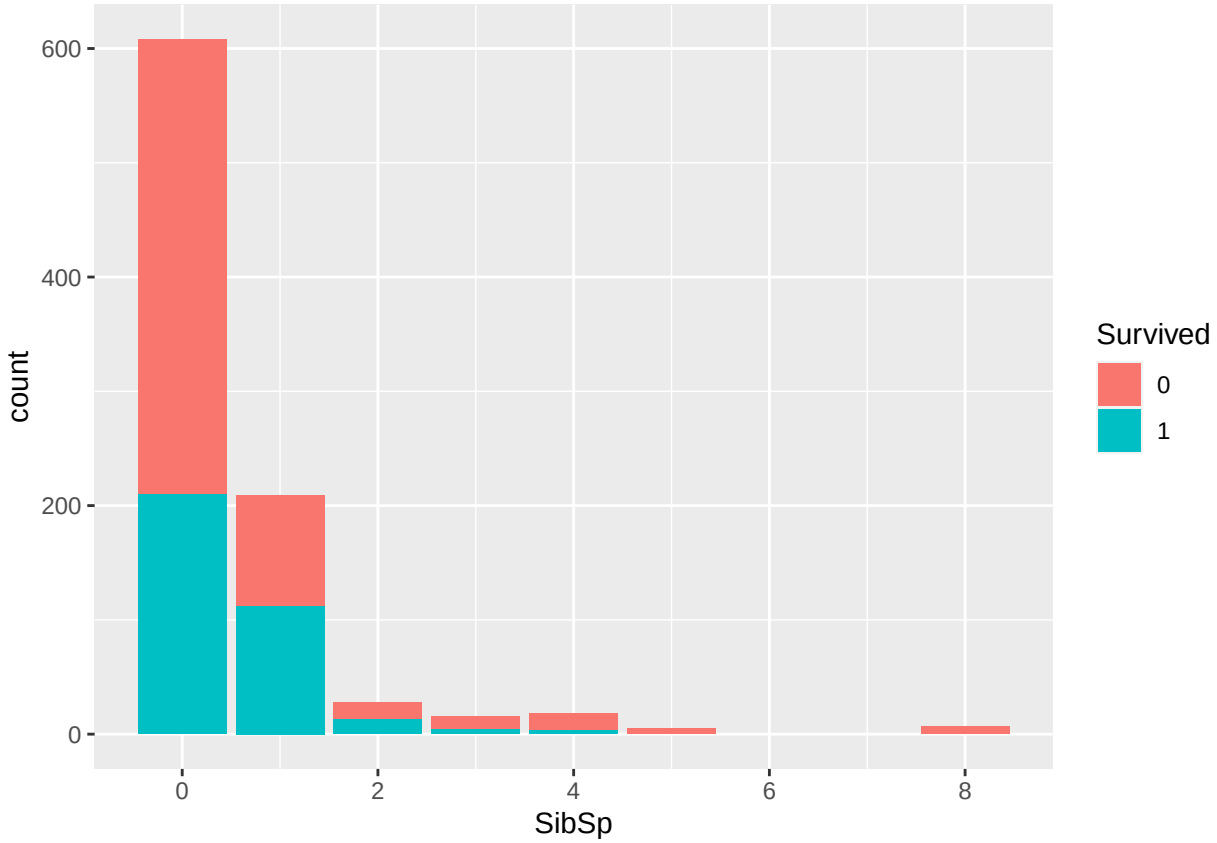
```
PSexo<-ggplot(train, aes(x=Sex, fill=Survived)) + geom_bar()
PSexo
```



El sexo también impacta fuertemente sobre el índice de supervivencia, teniendo las mujeres más posibilidades de no morir.

Hermanos/hermanas, esposos/esposas vs survived

```
ggplot(train, aes(x=SibSp, fill=Survived)) + geom_bar()
```



No se aprecia diferencias significativas entre los diferentes elementos de esta agrupación, entendiéndosa que esta variable no afecta a la tasa de supervivencia.

Normalidad y homogeneidad de la varianza

Para verficar la suposición de la normalidad, utilizamos el test de Shapiro-Wilk, considerado uno de los métodos más potentes. Para el estudio de la homocedasticidad usamos el test de Fligner-Killeen, ya que se ha detectado previamente que los datos no cumplen con la condición de normalidad.

Edad

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

```
##
##  Shapiro-Wilk normality test
##
## data:  train$Age
## W = 0.97807, p-value = 2.557e-10
```

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

```
##
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age.y by Survived
## Fligner-Killeen:med chi-squared = 3.4087, df = 1, p-value = 0.06485
```

Normalidad: en el caso de la edad, p-value es menor que 0.05, por lo que se acepta la hipótesis nula del test, con lo que se concluye que la variable Age no sigue una distribución normal. Homocedasticidad: p-value es superior a 0,05, se acepta la hipótesis nula de homocedasticidad, como conclusión, la variable Edad no presenta varianzas estadísticamente diferentes para los diferentes grupos de Supervivencia.

Padres/madres, hijos e hijas

```
shapiro.test(train$Parch)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train$Parch
## W = 0.53281, p-value < 2.2e-16
```

Normalidad: en este caso, para Parch, p-value es menor que 0.05, por lo que se acepta la hipótesis nula del test, con lo que se concluye que la variable Age no sigue una distribución normal. Homocedasticidad: p-value es inferior a 0,05, no se acepta la hipótesis nula de homocedasticidad, como conclusión, la variable Parch presenta varianzas estadísticamente diferentes para los diferentes grupos de Supervivencia.

Tarifa del pasajero

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

```
##
##  Shapiro-Wilk normality test
##
## data:  train$Fare
## W = 0.51843, p-value < 2.2e-16
```

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 93.026, df = 1, p-value < 2.2e-16
```

Normalidad: en el caso de la variable Tarifa, p-value es menor que 0.05, por lo que se acepta la hipótesis nula del test, con lo que se concluye que la variable Age no sigue una distribución normal. Homocedasticidad: p-value es inferior a 0,05, no se acepta la hipótesis nula de homocedasticidad, como conclusión, la variable Fare presenta varianzas estadísticamente diferentes para los diferentes grupos de Supervivencia.

Hermanos/hermanas, esposos/esposas

```
shapiro.test(train$SibSp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  train$SibSp
## W = 0.51297, p-value < 2.2e-16
```

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

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  SibSp by Survived
## Fligner-Killeen:med chi-squared = 1.2514, df = 1, p-value = 0.2633
```

Normalidad: Para la variable SibSp, p-value es menor que 0.05, por lo que se acepta la hipótesis nula del test, con lo que se concluye que la variable Age no sigue una distribución normal. Homocedasticidad: p-value es superior a 0,05, se acepta la hipótesis nula de homocedasticidad, como conclusión, la variable SibSp no presenta varianzas estadísticamente diferentes para los diferentes grupos de Supervivencia.

Comparación de grupos

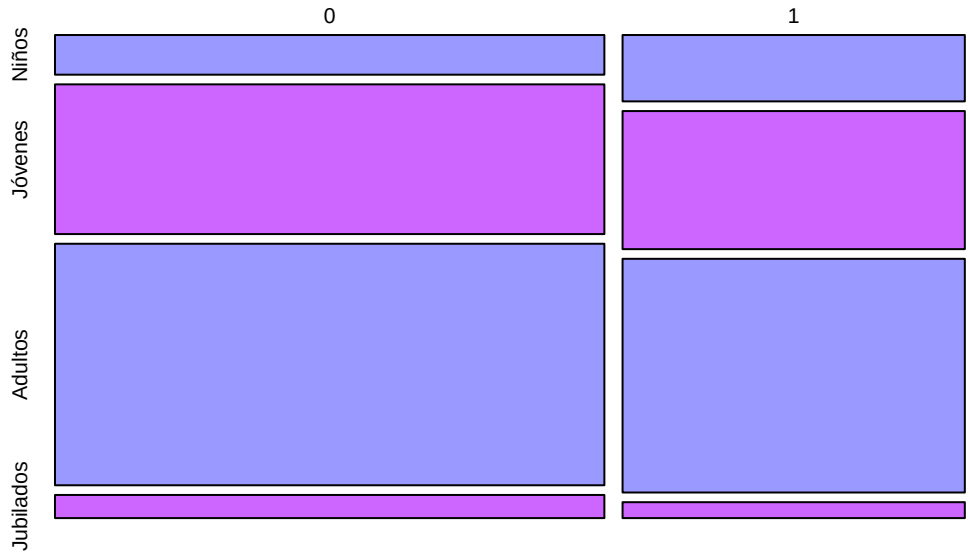
Se van a comparar distintos grupos: Grupo de Edad, Clase, Sexo, Tarifa, hermanos/hermanas, esposos/esposas, padres/madres, hijos/hijas y Supervivencia

Primero se realizan los contrastes de Hipótesis entre los grupos con Survived

Grupo de Edad vs Supervivencia

Por el tipo de variables, se puede utilizar el test chi-cuadrado:

```
temporal<-table(train$Survived, train$GrupoEdad)
plot(temporal, col=c("#9999FF","#CC66FF"), main="GrupoEdad vs Supervivientes")
```




```
chisq.test(temporal)
```

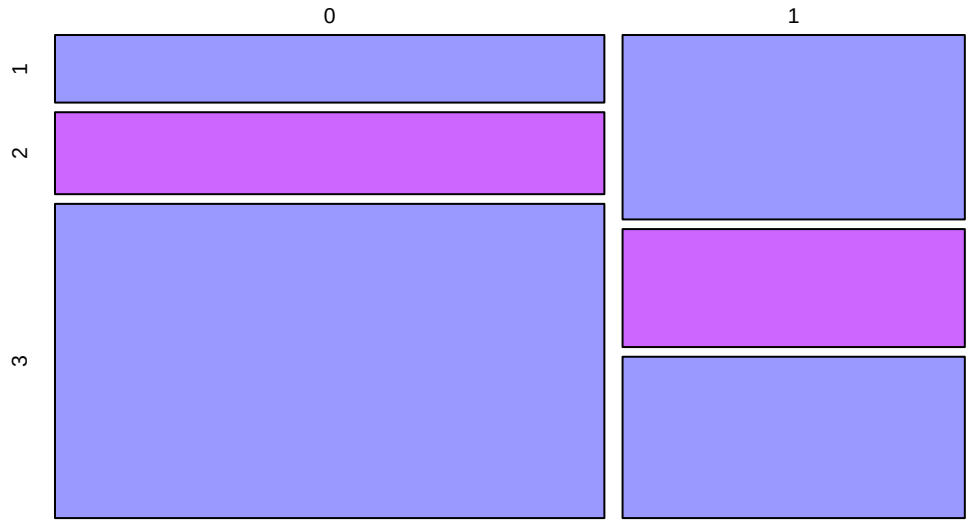
```
##
##  Pearson's Chi-squared test
##
## data:  temporal
## X-squared = 8.3566, df = 3, p-value = 0.03919
```

Dependencia: p-value es inferior a $< 0{,}05$, se rechaza la hipótesis nula de independencia con lo que la supervivencia depende del grupo de edad.

Clase vs Supervivencia

Por el tipo de variables, también se puede utilizar el test chi-cuadrado:

```
temporal<-table(train$Survived, train$Pclass)
plot(temporal, col=c("#9999FF","#CC66FF"), main="Clase vs Supervivientes")
```




```
chisq.test(temporal)
```

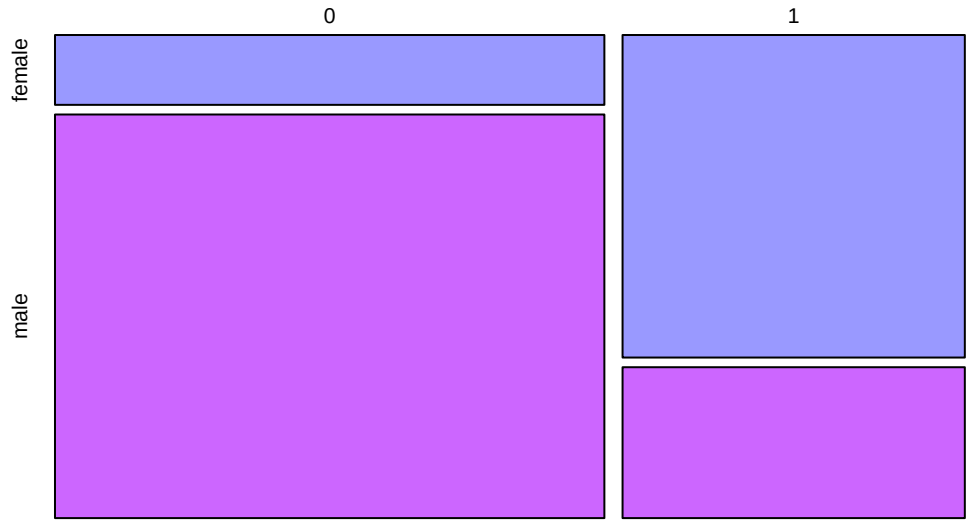
```
##
##  Pearson's Chi-squared test
##
## data:  temporal
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

Dependencia: p-value es inferior a $< 0{,}05$, se rechaza la hipótesis nula de independencia con lo que la supervivencia depende de la clase a la que pertenece el ticket.

Sexo vs Supervivencia

Por el tipo de variables, se puede utilizar el test chi-cuadrado:

```
temporal2<-table(train$Survived, train$Sex)
plot(temporal2, col=c("#9999FF","#CC66FF"), main="Sexo vs Supervivencia")
```




```
chisq.test(temporal2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  temporal2
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Dependencia: p-value es inferior a $< 0{,}05$, se rechaza la hipótesis nula de independencia con lo que la supervivencia depende del sexo del pasajero.

Tarifa vs Supervivencia

Se va a realizar una regresión lineal para aproximar la relación de dependencia lineal entre las dos variables, mediante la función lm().

datFare=lm(Survived ~ Fare, data=train) summary(datFare)

Se aprecia un R-squared bajo, con lo que las variables no se correlacionan.

Padres/madres, hijos e hijas vs Supervivencia

Se va a realizar una regresión lineal para aproximar la relación de dependencia lineal entre las dos variables, mediante la función lm().

datParch=lm(Survived ~ Parch, data=train) summary(datParch)

A continuación, se va a ejecutar el test de ANOVA, para confirmar que la diferencia con y sin la variable, no es significativa.

```
tieneParch <- glm(Survived ~ Parch, family = binomial(link='logit'), data = train)
summary(tieneParch)
```

```
##
## Call:
## glm(formula = Survived ~ Parch, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4705  -0.9533  -0.9533   1.4195   1.4195
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.55305    0.07689  -7.192 6.37e-13 ***
## Parch        0.20332    0.08462   2.403   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance: 1180.8  on 889  degrees of freedom
## AIC: 1184.8
##
## Number of Fisher Scoring iterations: 4
```

```
anova(tieneParch, test="Chisq")
```

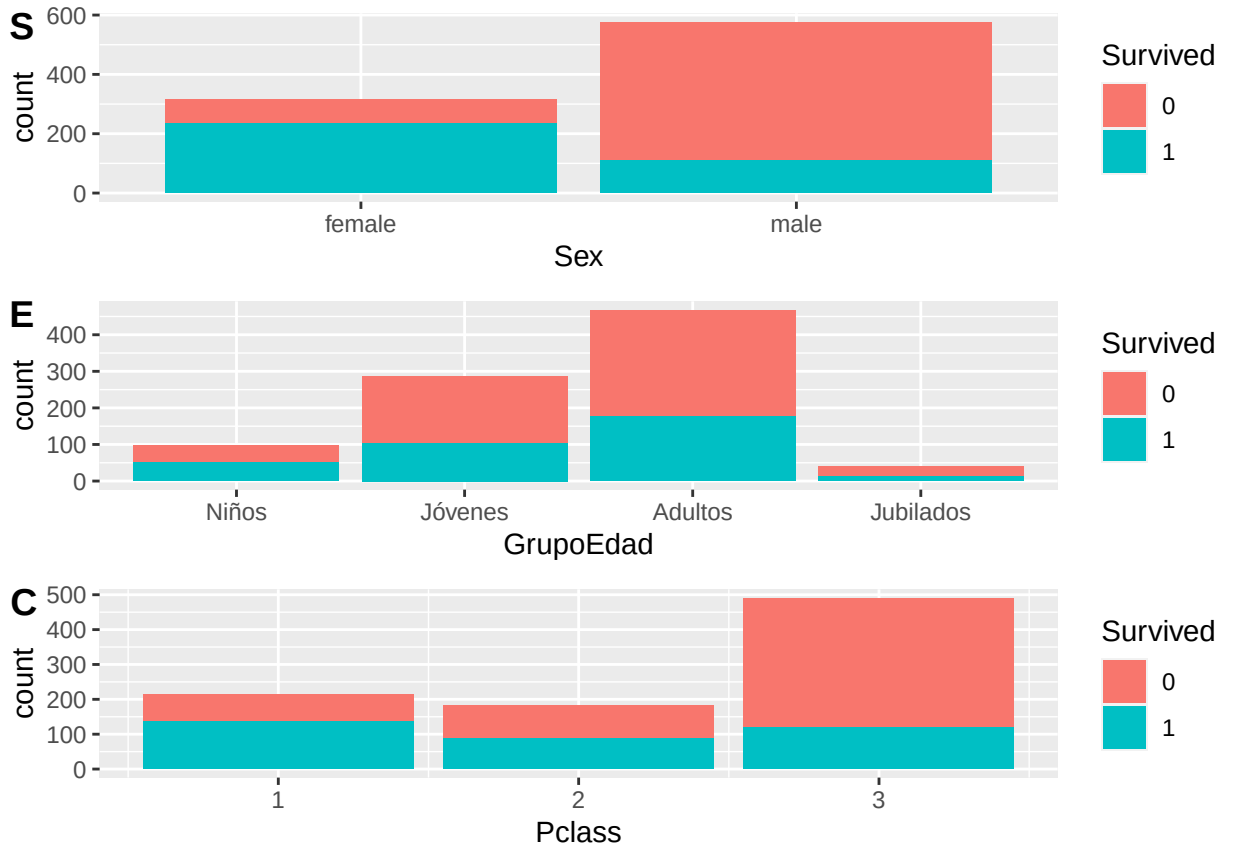
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    890     1186.7
## Parch  1   5.8135       889     1180.8   0.0159 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede apreciar la desviación de residuales es prácticamente la misma sin la variable que con la variable (1186.7 vs 1180.8). La variable Parch no depende de la variable Survived.

Resultados y Conclusiones

Del estudio previo resulta que las variables Sex, Pclass y Age son las que tienen mayor relación con Survived.

```
ggarrange(PSexo, PGedad, PClase, labels = c("S", "E", "C"), ncol = 1, nrow = 3)
```



Para evaluar la posibilidad de supervivencia, se van a crear diferentes modelos de predicción, para valorarlos a continuación. En primer lugar se van a dividir de nuevo los datos para realizar el análisis

```
set.seed(345)
```

Y se crean las diferentes regresiones:

Survived vs PClass + Sex + Age Survived vs PClass + Sex Survived vs PClass

```
M0 <- glm( formula = Survived ~ Pclass + Sex + Age.y, data = train, family = binomial)
summary(M0)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age.y, family = binomial,
##     data = train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6544  -0.6384  -0.4185   0.6310   2.4355
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.77876    0.46414  10.296  < 2e-16 ***
## Pclass      -1.22067    0.12518  -9.751  < 2e-16 ***
## Sexmale     -2.56213    0.18617 -13.762  < 2e-16 ***
## Age.y       -0.03261    0.00735  -4.437 9.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  806.5  on 887  degrees of freedom
## AIC: 814.5
##
## Number of Fisher Scoring iterations: 5
```

```
M1 <- glm( formula = Survived ~ Pclass+ Sex, data = train, family = binomial)
summary(M1)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2030  -0.7036  -0.4519   0.6719   2.1599
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.2946     0.2974  11.077   <2e-16 ***
## Pclass       -0.9606     0.1061  -9.057   <2e-16 ***
## Sexmale      -2.6434     0.1838 -14.380   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  827.2  on 888  degrees of freedom
## AIC: 833.2
##
## Number of Fisher Scoring iterations: 4
```

```
M3 <- glm( formula = Survived ~ Pclass, data = train, family = binomial)
summary(M3)
```

```
##
## Call:
```

```
## glm(formula = Survived ~ Pclass, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4390  -0.7569  -0.7569   0.9367   1.6673
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.44679    0.20743   6.975 3.06e-12 ***
## Pclass      -0.85011    0.08715  -9.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance: 1084.4  on 889  degrees of freedom
## AIC: 1088.4
##
## Number of Fisher Scoring iterations: 4
```

Con el Dato AIC, llegamos a la conclusión de que el primer modelo, con las tres variables, es el mejor, con lo que la supervivencia de los pasajeros depende del sexo, edad y clase.