Práctica 2 - Limpieza y análisis de datos

Maria Dolores Moyano Guerrero y Victor Cancer Castillo

25 de Mayo de 2022

Contents

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

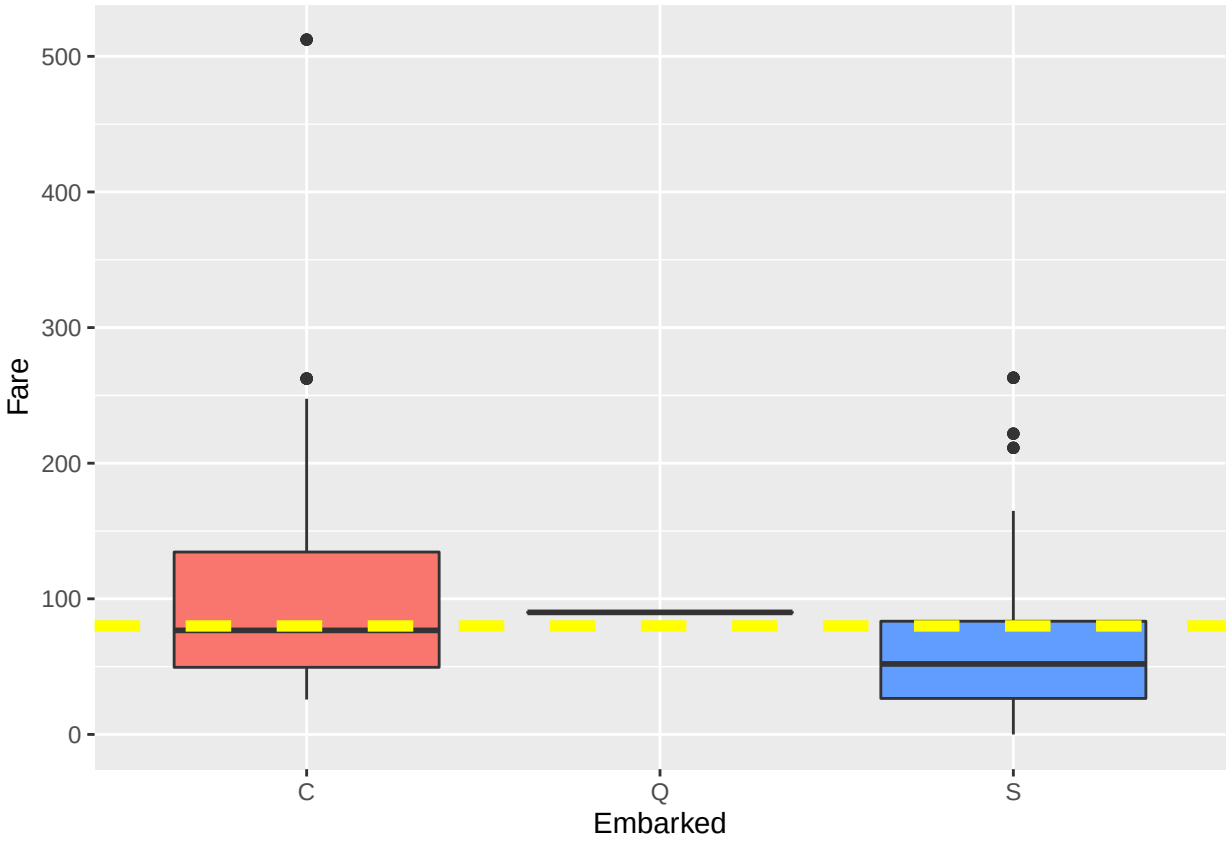
```
full[full$Embarked == "",]
```

```
##      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:

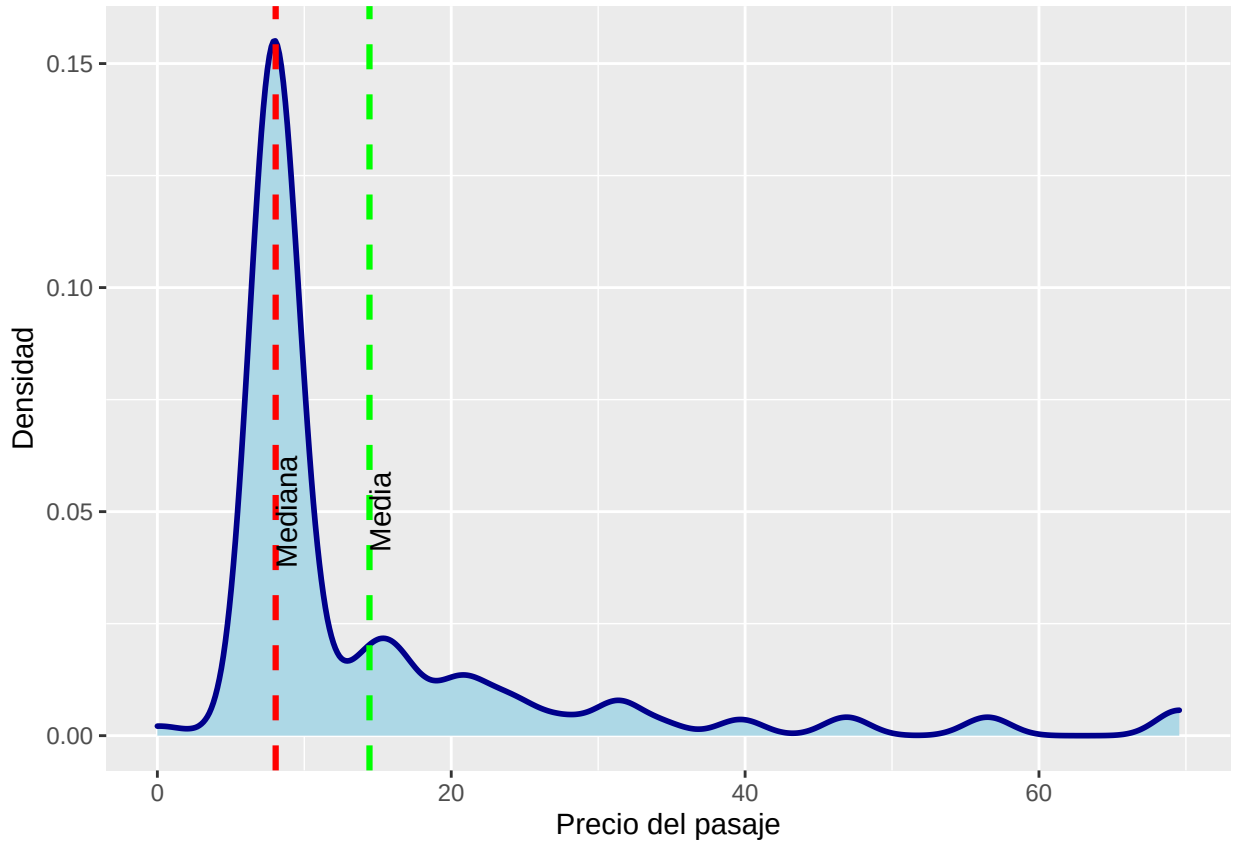
```
full[is.na(full$Fare),]
```

```
##     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" ,]$
  geom_vline(xintercept = mean(full[!is.na(full$Fare) & full$Pclass == "3" & full$Embarked == "S" ,]$Fa
```

```
  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.

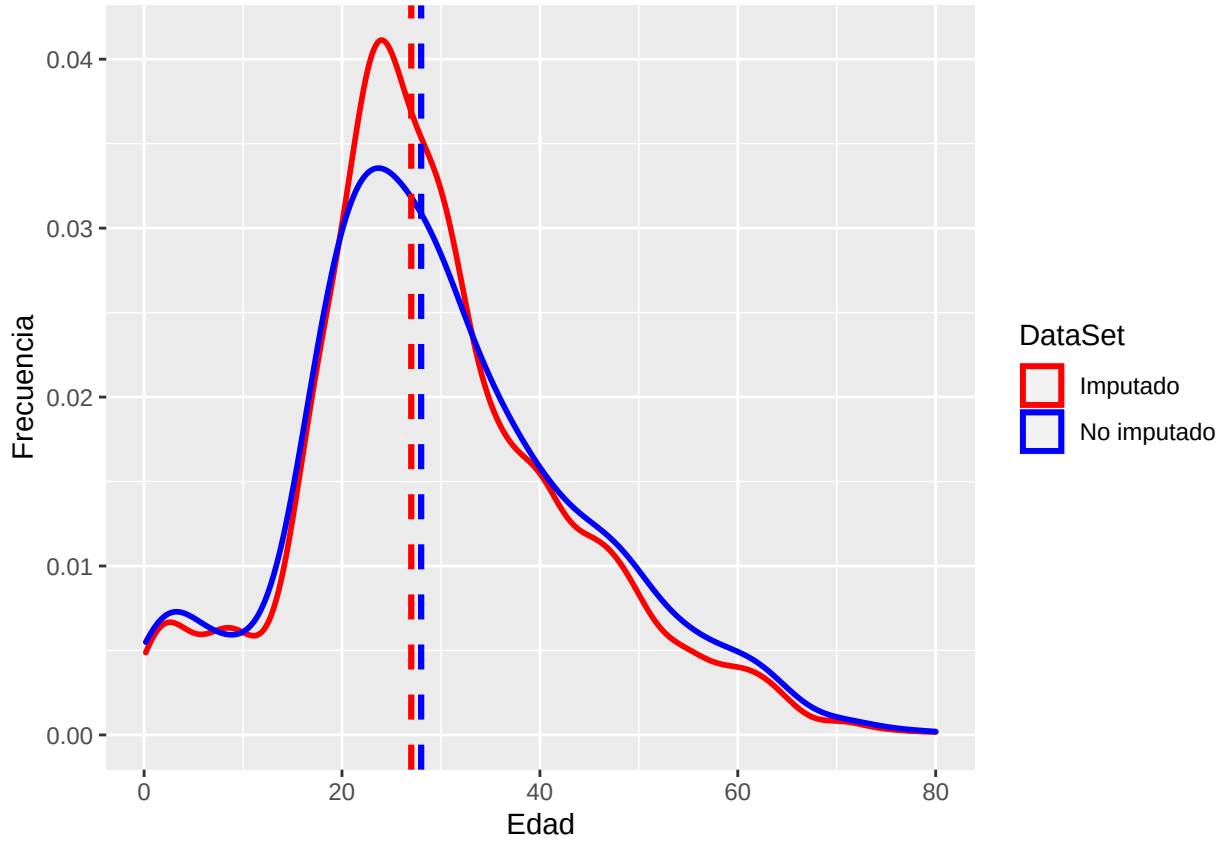
Las variables que tendremos en cuenta en esta imputación serán:

- 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.

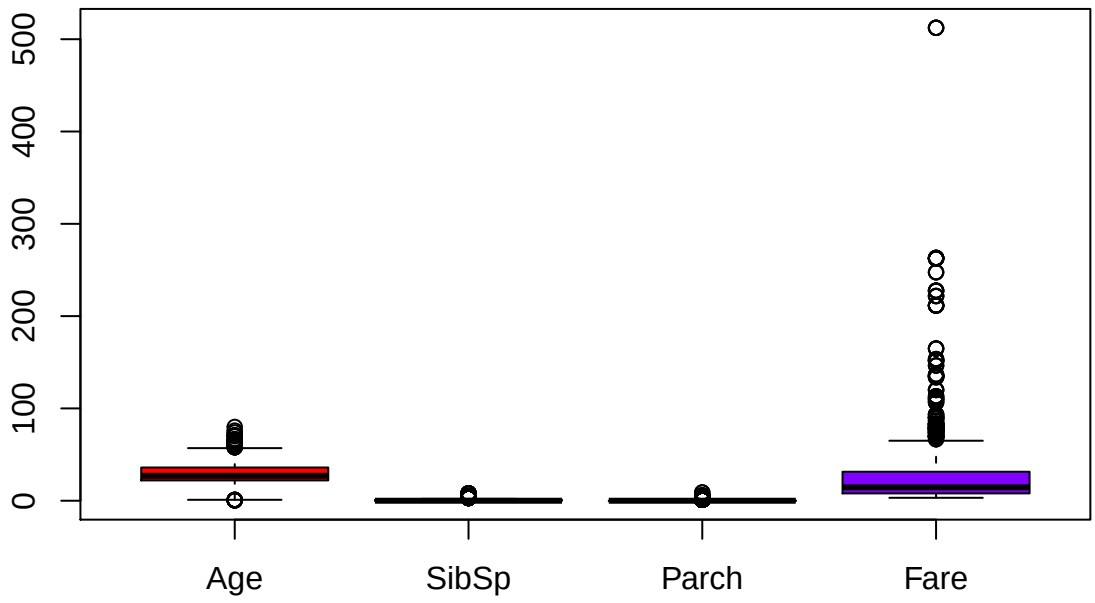
Por lo tanto pasamos ahora a imputar estos valores en los datasets que estamos ahora gestionando:

```
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 <- 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 <- 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)))
```



Revisando los valores extremos de edad vemos que son valores válidos

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

```
## [1] 0.17
```

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

```
## [1] 80
```

Para el fare (tarifa del pasajero) encontramos:

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

```
## [1] 66.6
```

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

```
## [1] 512.3292
```

Se ha buscado el rango de precios de los billetes (https://www.20minutos.es/noticia/1365526/0/titanic/hundimiento/aniversario/), y los precios máximosy mínimos 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

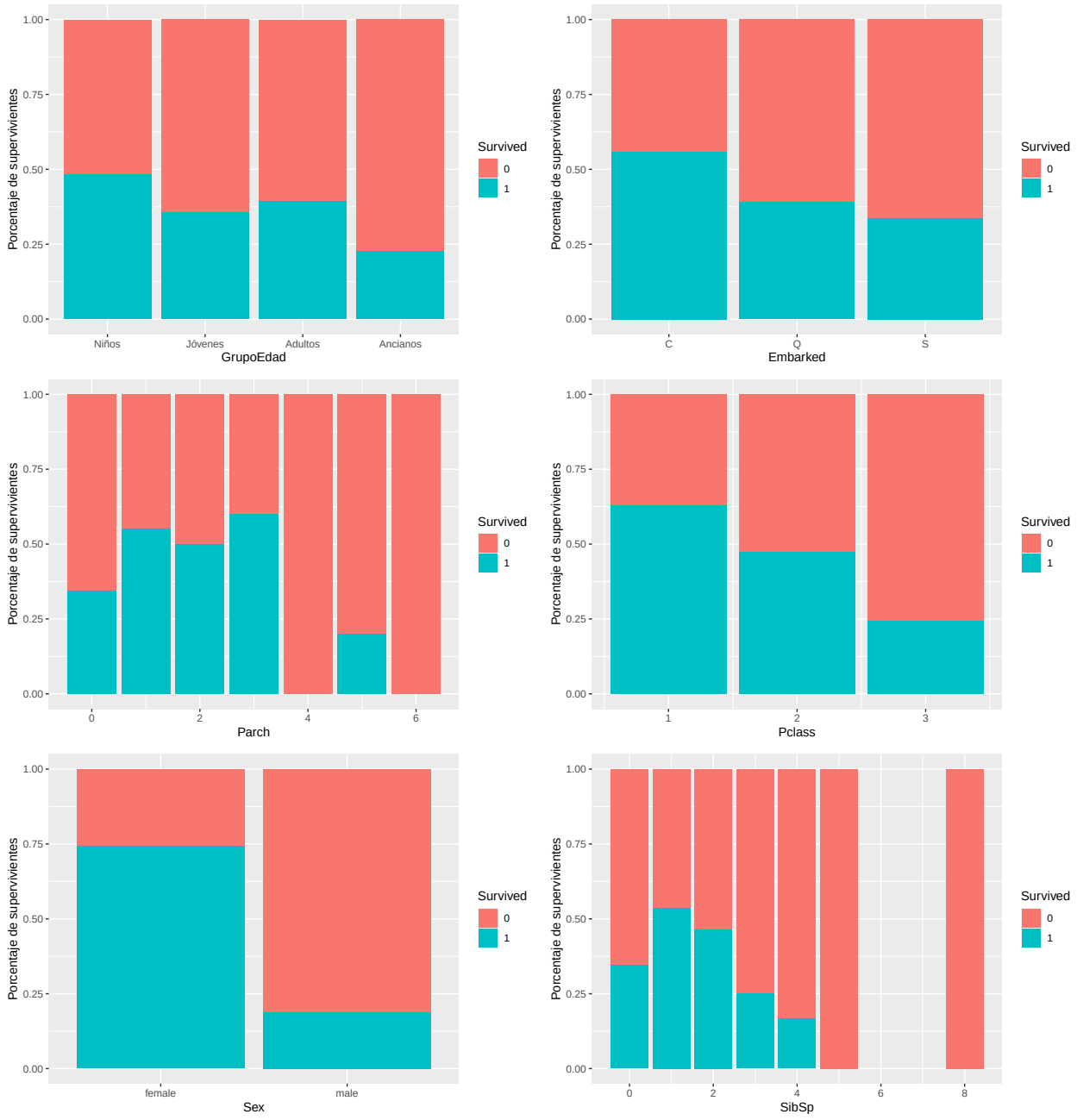
Los grupos seleccionados serán los siguientes, para estudiar su relación con survived:

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*: influencia del sexo del pasajero en la supervivencia. *SibSp* y *Parch*: influencia del número de hermanos/hermanas, esposos/esposas a bordo del Titanic en la supervivencia.

Vamos a hacer un primer análisis descriptivo de cual podría ser la relacion entre estas variables y la probabilida de supervivencia de los pasajeros

```
#Edad
train$GrupoEdad <- cut(train$Age, breaks = c(0,16,30,60,100), labels = c("Niños","Jóvenes","Adultos","A
train$Survived <- as.factor(train$Survived)
PGedad<-ggplot(train, aes(x=GrupoEdad, fill=Survived))  + geom_bar(position='fill') + ylab('Porcentaje

Pembarked <-ggplot(train, aes(x=Embarked, fill=Survived))  + geom_bar(position='fill') + ylab('Porcenta

Pparch <-ggplot(train, aes(x=Parch, fill=Survived))  + geom_bar(position='fill') + ylab('Porcentaje de

PClase<-ggplot(train, aes(x=Pclass, fill=Survived)) + geom_bar(position='fill') + ylab('Porcentaje de su

PSexo<-ggplot(train, aes(x=Sex, fill=Survived))  + geom_bar(position='fill') + ylab('Porcentaje de supe

PSibSp <- ggplot(train, aes(x=SibSp, fill=Survived))  + geom_bar(position='fill') + ylab('Porcentaje de
```

```
PGedad
Pembarked
Pparch
PClase
PSexo
PSibSp
```



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

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

Parch: Parece ser que los pasajeros con 1 a 3 padres/hijos tenian más probabilidades de sobrevivir.

Class: La clase es una variable que impacta fuertemente sobre la tasa de supervivencia, siendo la tercera clase la más afectada por el accidente.

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

SibSp: Parece que tener algún familiar puede aumentar tu probabilidad de sobrevivir, aunque ésta desciende conforme se tienen más familiares.

Normalidad y homogeneidad de la varianza

Normalidad

Para verficar la suposición de la normalidad, utilizamos el test de Shapiro-Wilk, considerado uno de los métodos más potentes, en las variables númericas

Variable	p-value Shapiro Test	Normalidad
Age	\ensuremath{2.5566505\times 10^{-10}}	Distribución normal
Parch	\ensuremath{2.3866223\times 10^{-43}}	Distribución normal
Fare	\ensuremath{8.4686478\times 10^{-44}}	Distribución normal
SibSp	\ensuremath{5.7508309\times 10^{-44}}	Distribución normal
Fare	\ensuremath{8.4686478\times 10^{-44}}	Distribución normal

Se encuentra en todos los casos que el p-value es menor a 0.05, con lo que todos siguen una distribución normal.

Homocedasticidad

Para el estudio de la homocedasticidad usamos el estadístico F, que se puede aplicar con la función *var.test()*. Lo aplicaremos para unos grupos a modo de ejemplo

```
var.test(x=train[train$Embarked=='S','Fare'],y=train[train$Embarked=='C','Fare'])
```

```
##
##  F test to compare two variances
##
## data:  train[train$Embarked == "S", "Fare"] and train[train$Embarked == "C", "Fare"]
## F = 0.18366, num df = 643, denom df = 169, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1432147 0.2315569
## sample estimates:
## ratio of variances
##          0.1836642
```

Al comparar los precios de los billetes de los puertos de embarque S y C encontramos que hay una diferencia significativa entre las varianzas de los dos grupos.

Podemos aplicar este mismo test para tratar de encontrar si hay homogeneidad en la varianza para los sexos en la variable de edad

```
var.test(x=train[train$Sex=='male','Age'],y=train[train$Sex=='female','Age'])
```

```
##
##  F test to compare two variances
##
## data:  train[train$Sex == "male", "Age"] and train[train$Sex == "female", "Age"]
## F = 1.0042, num df = 576, denom df = 313, p-value = 0.9739
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8240269 1.2169029
```

```
## sample estimates:
## ratio of variances
##           1.004235
```

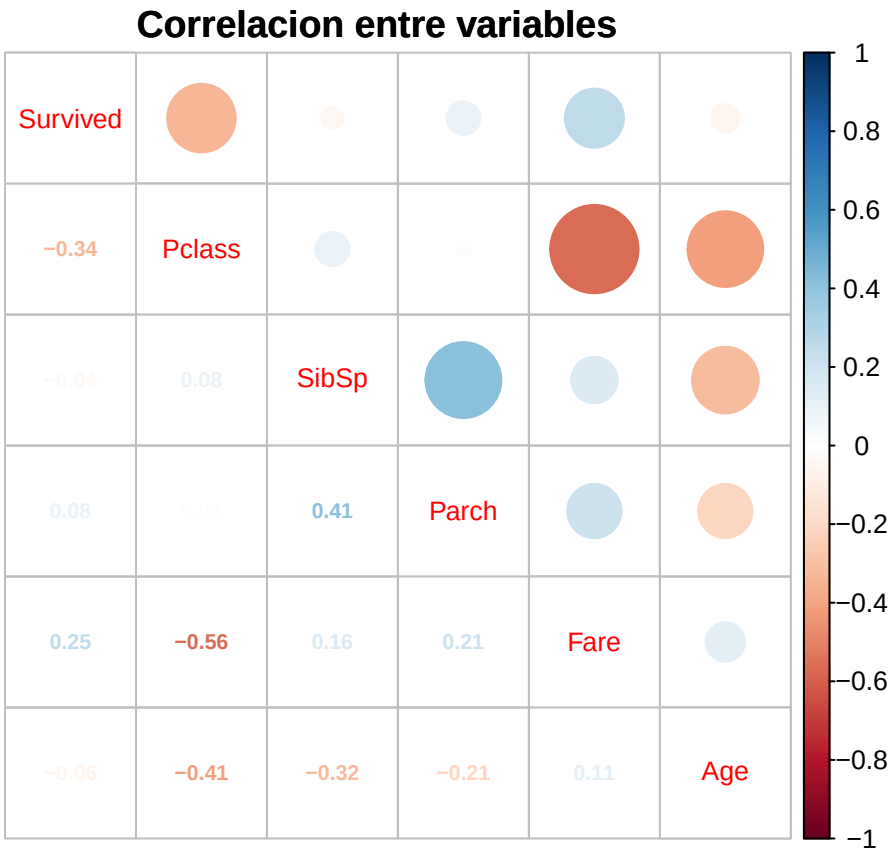
En este caso encontramos que las varianzas no muestran diferencias significativas entre sexos.

Comparación de grupos

Correlación entre variables

Nos interesa saber si hay posibles relaciones entre las variables que estamos teniendo en cuenta, por lo que haremos un calculo de la matriz de correlación para las variables númericas

```
cor_table <- cor(train[,c("Survived","Pclass","SibSp","Parch","Fare","Age")],use = "complete.obs")
corrplot.mixed(cor_table,upper="circle",number.cex=.7,tl.cex=.8, title="Correlacion entre variables", ma
```



Vemos que hay una clara relación entre la clase del pasaje y el precio de éste, como era de esperar. La edad también influye en qué tipo de pasaje se compra, así como su precio.

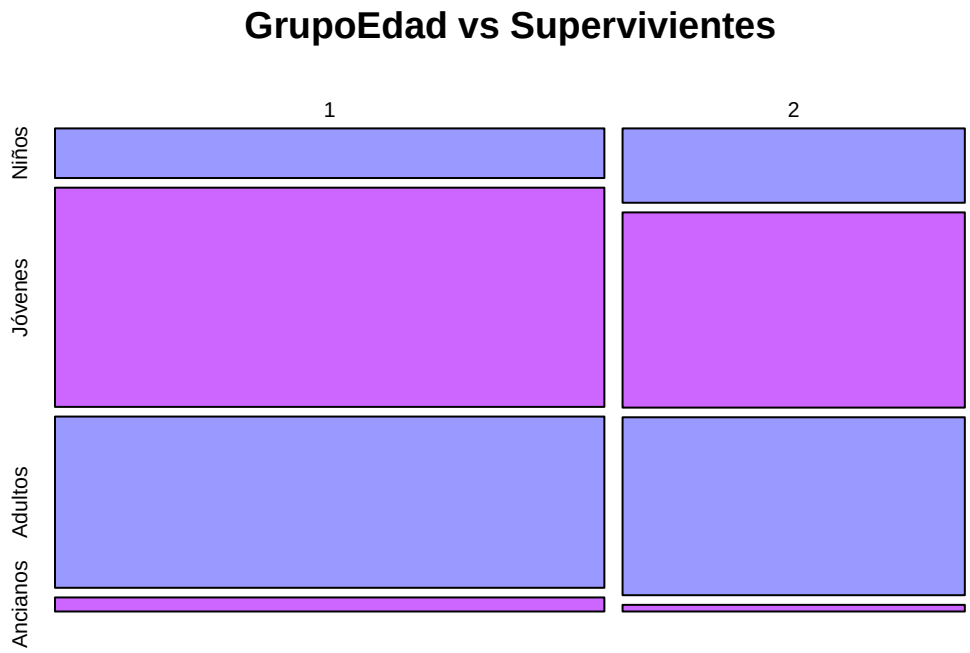
Otra relación que encontramos se da entre el numero de hijos-padres con hermanos-esposos, con un coeficiente de correlación de 0.38. De nuevo la edad vuelve a tener cierta importancia para estas variables.

Finalmente vemos que hay una clara relación entre la clase de pasaje y el la probabilidad de sobrevivir al accidente del Titanic.

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)
```

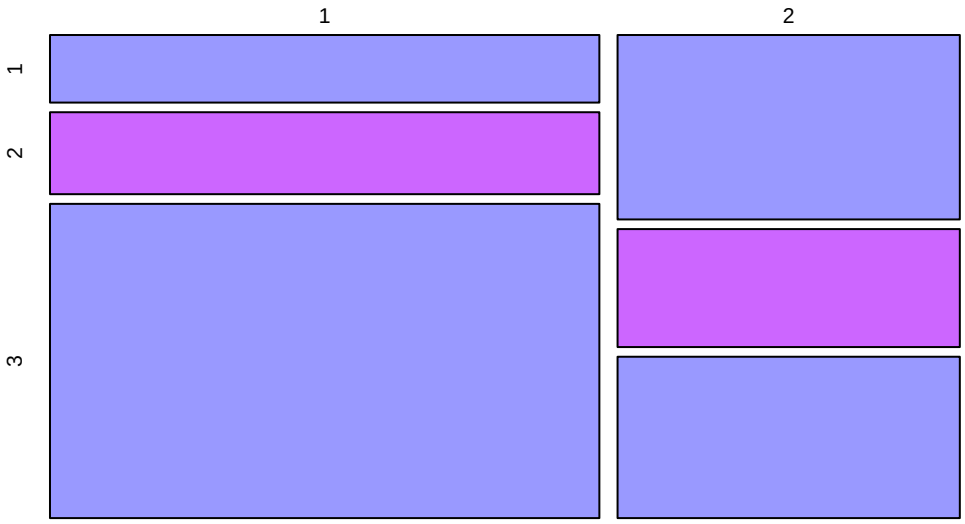
```
##
##  Pearson's Chi-squared test
##
## data:  temporal
## X-squared = 8.4735, df = 3, p-value = 0.03718
```

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)
```

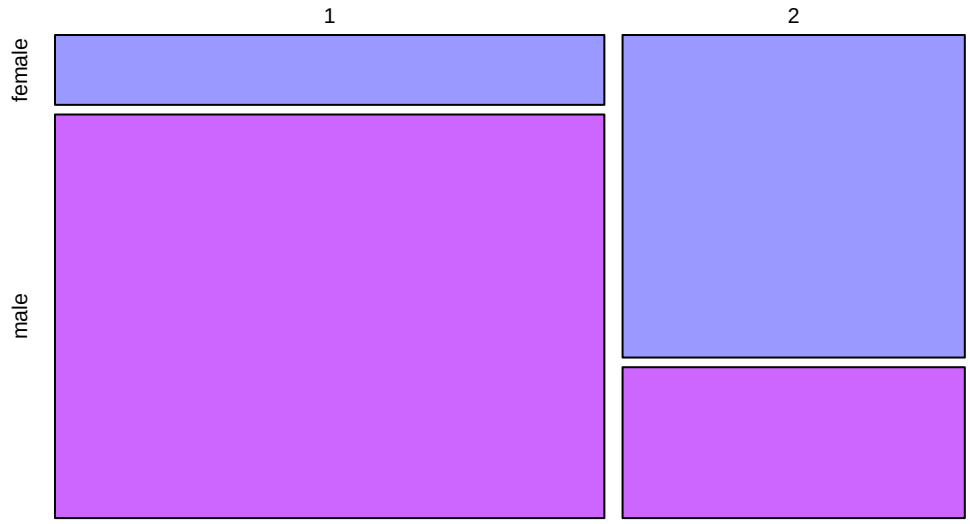
```
##
##  Pearson's Chi-squared test
##
## data:  temporal
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

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.

```r
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
```

```r
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
```

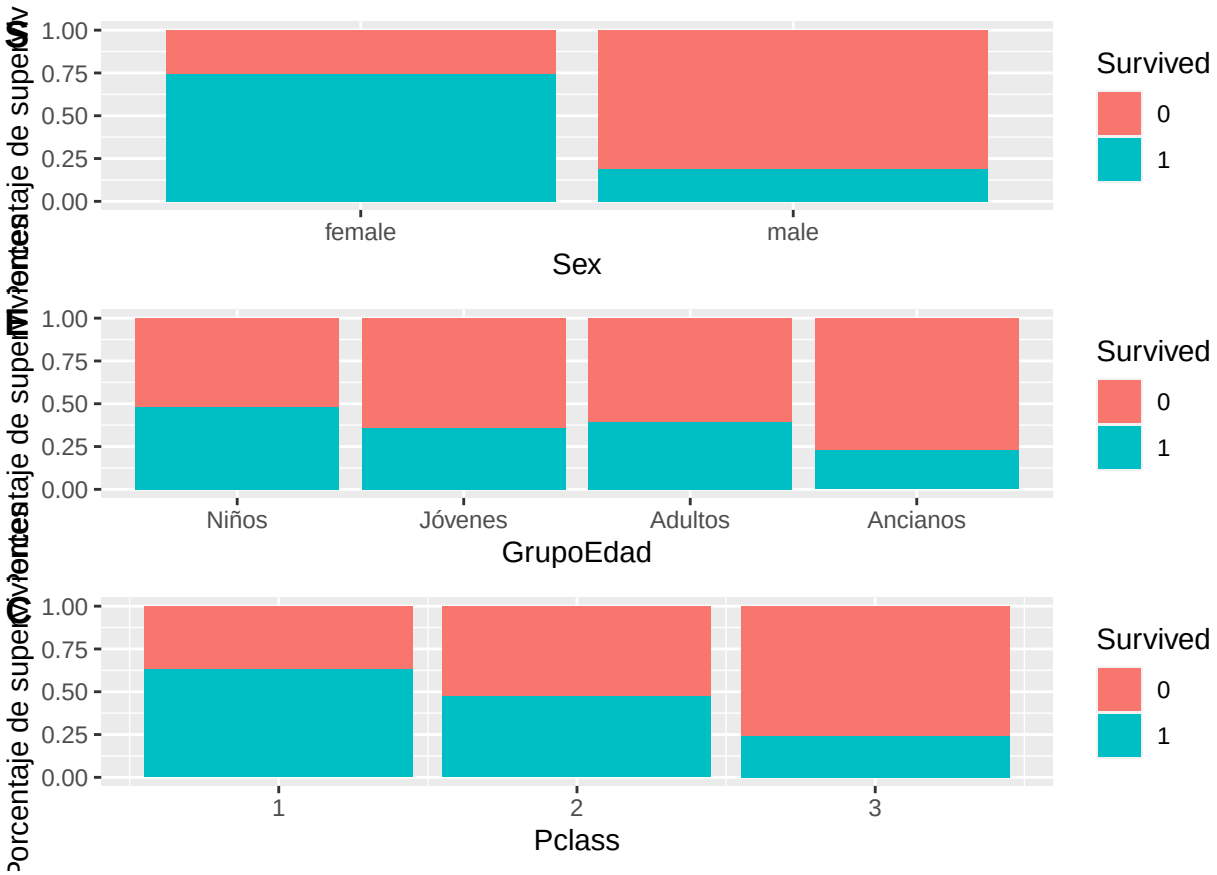
Se puede apreciar la desviación de residuales es prácticamente la misma sin la variable que con la variable (1186.7 vs 1180.8). La variable Parch no depende de la variable Survived.

Métodos de clasificación

Resultados y Conclusiones

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

```r
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, data = train, family = binomial)
summary(M0)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6451  -0.6491  -0.4166   0.6318   2.4383
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.532600   0.368655   9.582  < 2e-16 ***
## Pclass2     -1.155349   0.259883  -4.446 8.76e-06 ***
## Pclass3     -2.430315   0.253182  -9.599  < 2e-16 ***
## Sexmale     -2.561356   0.186301 -13.749  < 2e-16 ***
## Age         -0.032470   0.007362  -4.410 1.03e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  806.42  on 886  degrees of freedom
## AIC: 816.42
##
## 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
## -2.1877  -0.7312  -0.4476   0.6465   2.1681
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.2971     0.2190  10.490  < 2e-16 ***
## Pclass2      -0.8380     0.2447  -3.424 0.000618 ***
## Pclass3      -1.9055     0.2141  -8.898  < 2e-16 ***
## Sexmale      -2.6419     0.1841 -14.351  < 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: 1186.66  on 890  degrees of freedom
## Residual deviance:  826.89  on 887  degrees of freedom
## AIC: 834.89
##
## 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
## -1.4094  -0.7450  -0.7450   0.9619   1.6836
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5306     0.1409   3.766 0.000166 ***
## Pclass2      -0.6394     0.2041  -3.133 0.001731 **
## Pclass3      -1.6704     0.1759  -9.496  < 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: 1083.1  on 888  degrees of freedom
## AIC: 1089.1
##
## 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.