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

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

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*)

```r
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

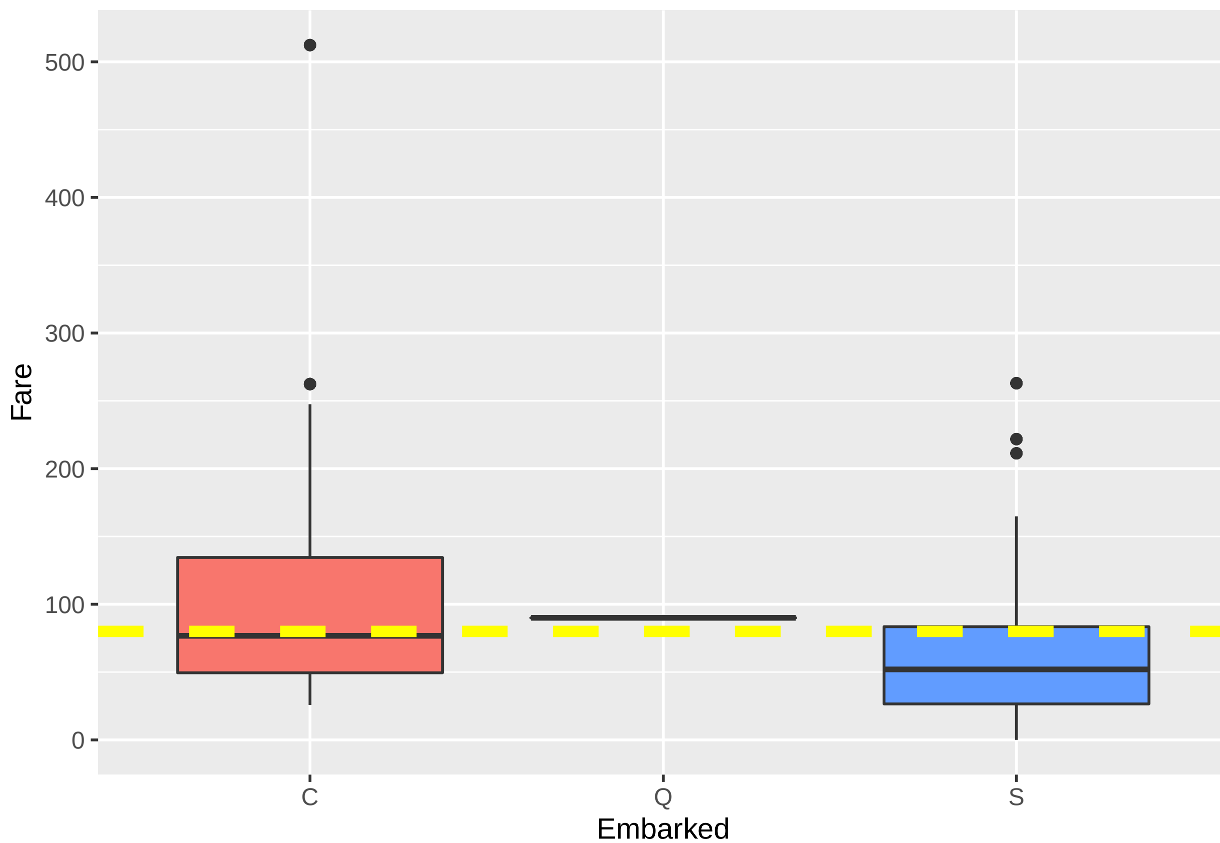
Vemos entre los valores de la columna Embarked del dataset de entrenamiento que hay dos valores vacíos

```r
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:

```r
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', l
```



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

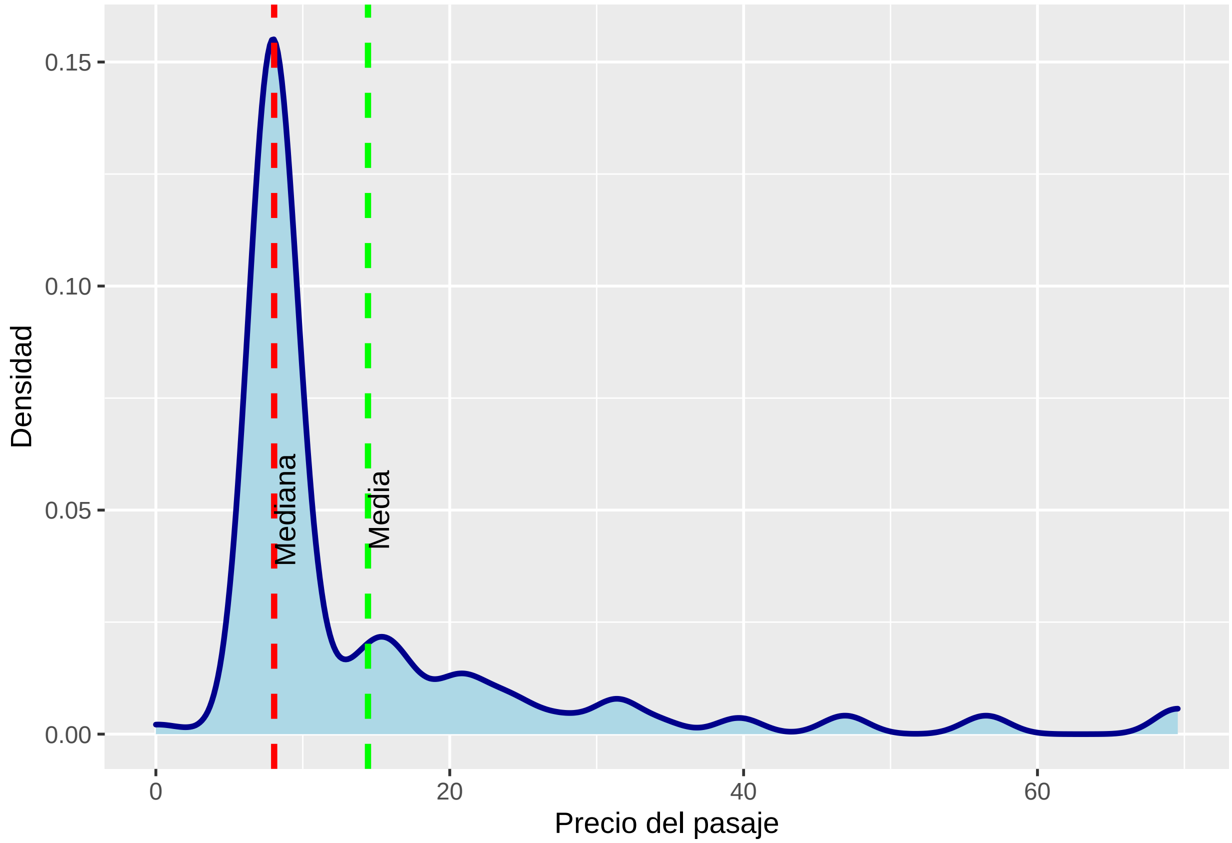
De las tarifas de los pasajes encontramos que tan solo hay un caso donde desconocemos el precio que se pagó:

```
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

```r
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

```r
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)

```r
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)
```

```r
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.

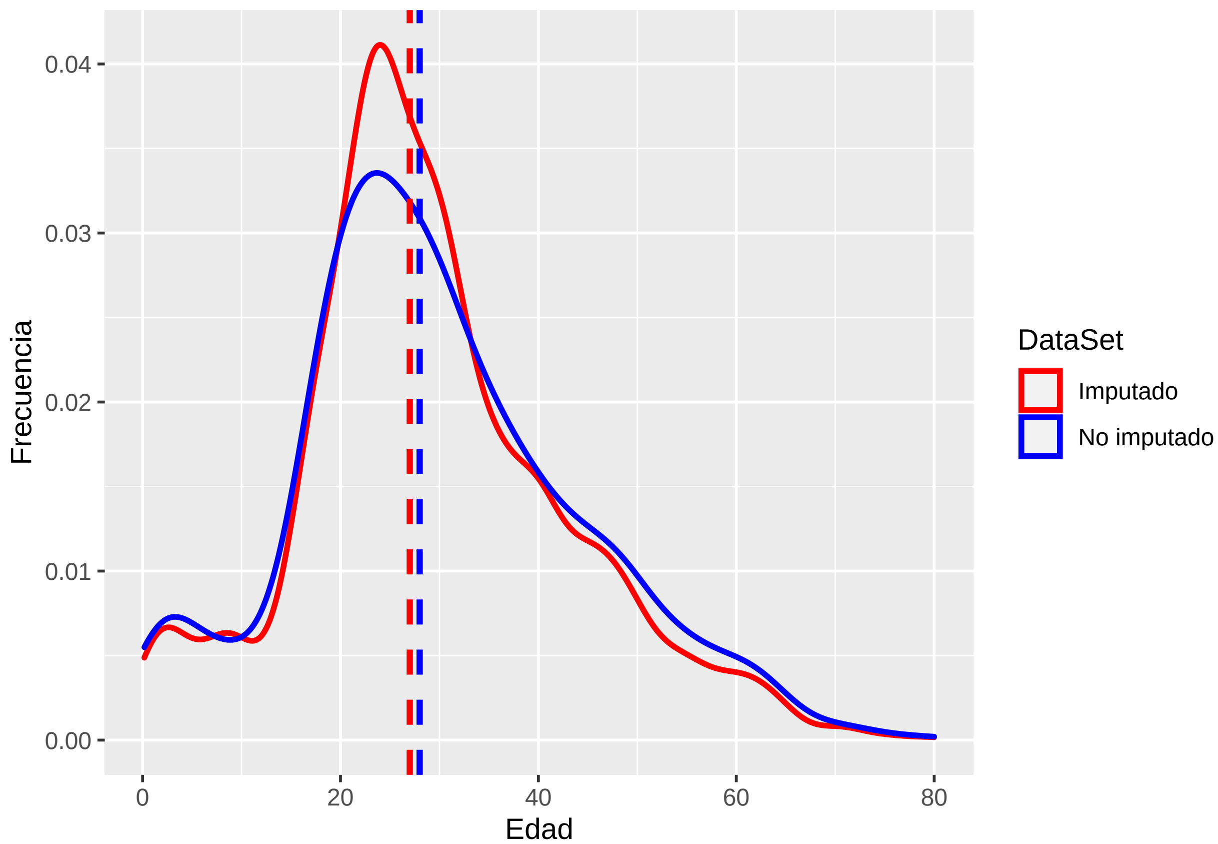
Las variables que tendremos en cuenta en esta imputación serán:

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

```r
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

```r
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.

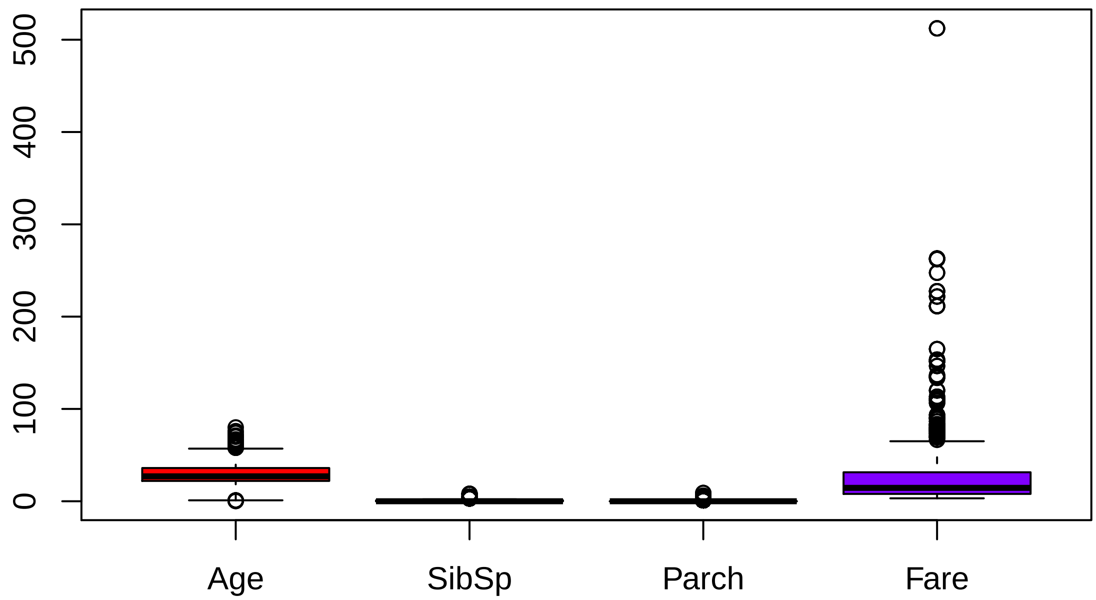
Por lo tanto pasamos ahora a imputar estos valores en los datasets que estamos ahora gestionando:

```
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

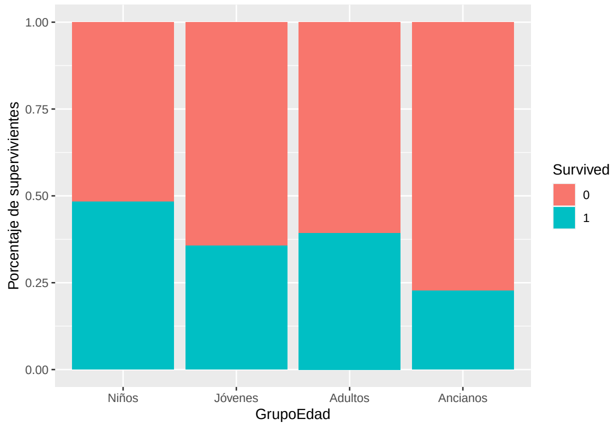
Los grupos seleccionados serán los siguientes, para estudiar su relación con survived:

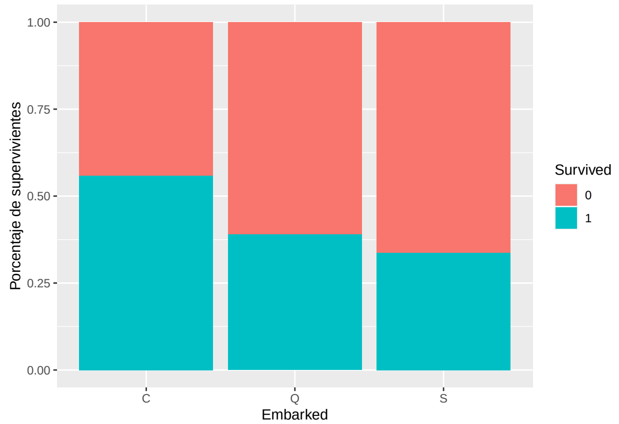
Age: se estudiará el efecto del rango de edad del pasajero en la supervivencia. *Embarked*: se analizará el efecto del puerto de embarque en la supervivencia. *Parch*: número de padres/madres, hijos/hijas a bordo del Titanic y su influencia. *Pclass*: se analizará la influencia de clase del pasajero. *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.

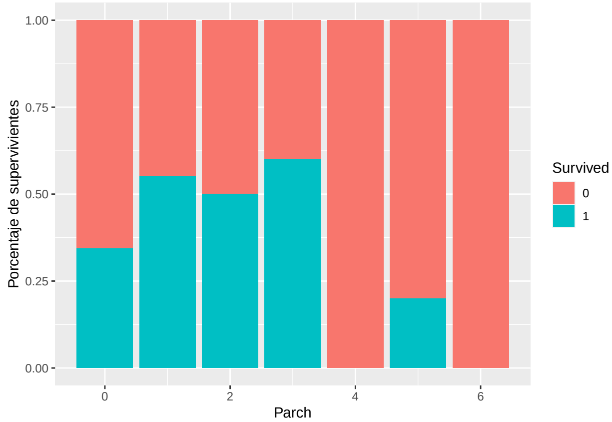
Vamos a hacer un primer análisis descriptivo de cual podría ser la relacion entre estas variables y la probabilida de supervivencia de los pasajeros

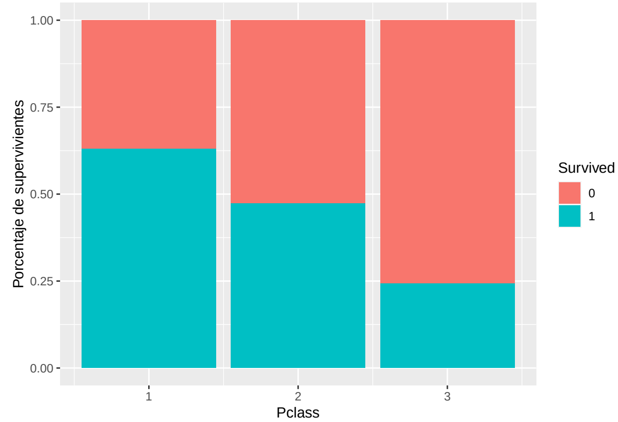
```
#Edad
train$GrupoEdad <- cut(train$Age, breaks = c(0,16,30,60,100), labels = c("Niños","Jóvenes","Adultos","An
train$Survived <- as.factor(train$Survived)
PGedad<-ggplot(train, aes(x=GrupoEdad, fill=Survived))  + geom_bar(position='fill') + ylab('Porcentaje d

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

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

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 super

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

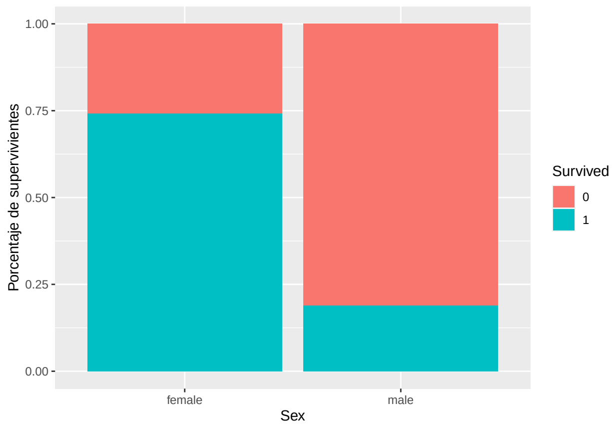
```
PGedad
Pembarked
Pparch
PClase
PSexo
PSibSp
```

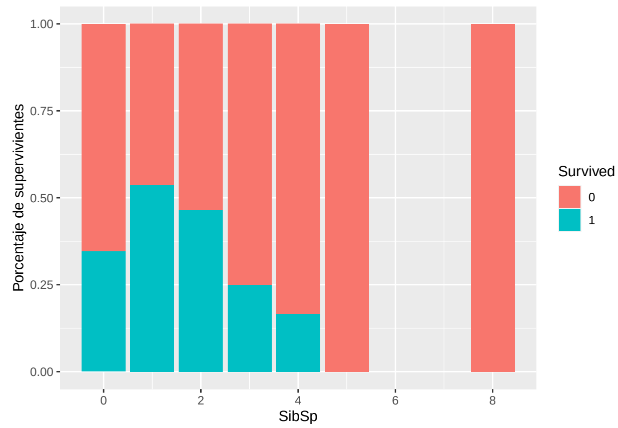












Age: Se aprecia que el porcentaje de supervivientes aumenta cuanto menor es la edad.

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	0	Distribución normal
Parch	0	Distribución normal
Fare	0	Distribución normal
SibSp	0	Distribución normal

Variable	p-value Shapiro Test	Normalidad
Fare	0	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
```

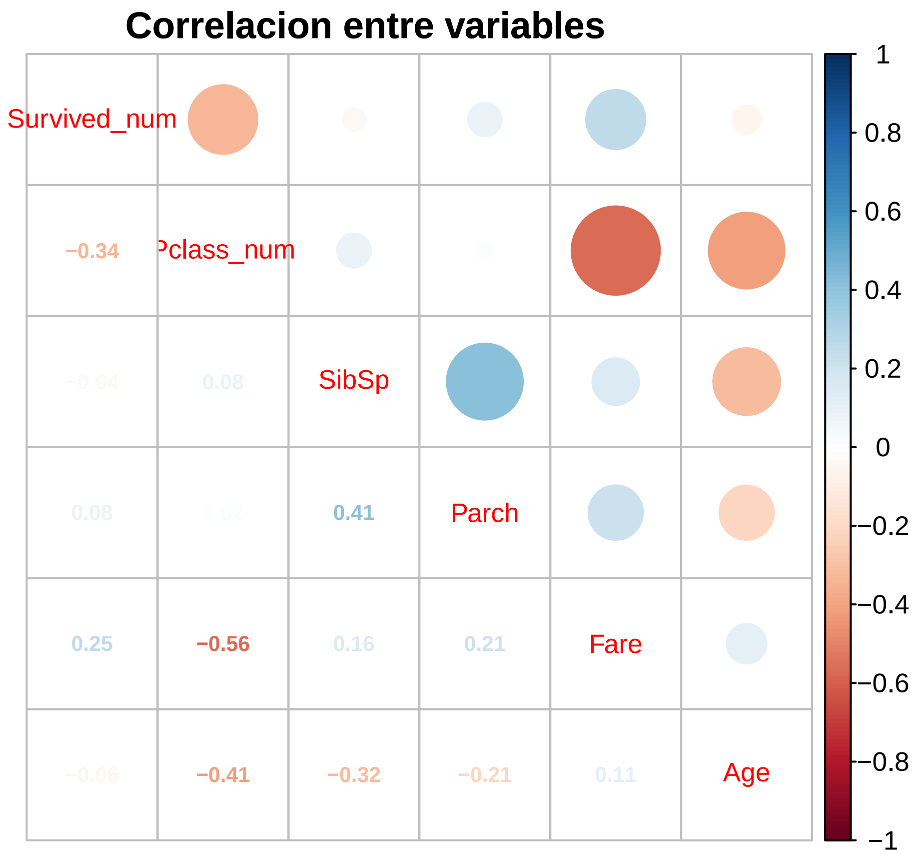
En este caso encontramos que las varianzas no muestran diferencias significativas entre sexos.

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_num","Pclass_num","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.

Contraste de hipotesis

Nos planteamos la siguiente pregunta: ¿es el porcentaje de fallecidos más alto en el el grupo de tercera clase que en el resto de clases?

La hipotesis nula y alternativa son entonces:

$$H_0 : p_{12} = p_3 \ ; \ H_1 : p_{12} > p_3$$

donde p_{12} es la proporción de supervivivientes de clase uno y dos y p_3 es la de tercera clase.

Puesto que nos hacemos la pregunta para inferir el valor en la población utilizando una muestra podemos asumir que la media va a seguir una distribución normal gracias al teorema del límite central. Podemos asumir lo mismo de la diferencia de las proporciones, $p_{12} - p_3$.

Sin embargo desconocemos la varianza de la población ni si las varianzas entre grupos son iguales, por lo que vamos a utilizar la diferencia de las proporciones :

$$z = \frac{(\hat{p}_{12} - \hat{p}_3) - (p_{12} - p_3)}{\sqrt{\frac{p_{12}(1-p_{12})}{n_{12}} - \frac{p_3(1-p_3)}{n_3}}} \sim N(0,1)$$

donde si se cumple la hipótesis nula, que es la hipótesis que queremos contrastar, tenemos $p_{12} = p_3 = p$ y por lo tanto el estadístico de contraste es

$$z = \frac{\hat{p}_{12} - \hat{p}_3}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_{12}} + \frac{1}{n_3}\right)}} \; ; \; \hat{p} = \frac{n_{12}\,\hat{p}_{12} + n_3\,\hat{p}_3}{n_{12} + n_3}$$

donde $\hat{p}$ es la estimación de la proporción poblacional común.

Vamos a definir nuestra propia función para llevar a cabo este análisis

```
test_prop <- function(x,NC){

  p1 <- sum(x$Survived == 1 & x$Pclass== 3) / sum(x$Pclass== 3)
  p2 <- sum(x$Survived == 1 & x$Pclass!=3) / sum(x$Pclass!= 3)

  #Podemos calcular p sin necesidad de utilizar la formula de arriba
  p <- sum(x$Survived == 1) / nrow(x)

  n1 <- sum(x$Pclass== 3)
  n2 <- sum(x$Pclass!= 3)

  zobs <- (p1-p2)/sqrt(p*(1-p)*((1/n1)+(1/n2)))

  zcrit <- qnorm(1-NC/100, lower.tail=TRUE)
  pobs <- pnorm(zobs,lower.tail=TRUE)

  round(c(zobs,zcrit,pobs,p1,p2),5)
}

test_prop(train,97)
```

```
## [1] -9.62078 -1.88079  0.00000  0.24236  0.55750
```

Por lo que concluimos con un nivel de confianza de 97% que un pasajero de tercera clase tiene más probabilidad de fallecer que si fuera de clase superior. Esto es debido a que z_{obs} es mucho mejor a z_{crit}, o dicho de otra manera, encontramos un p-value enormemente pequeño, lo cual nos permite descartar la hipótesis nula.

Regresión logística

Queremos crear un modelo de regresión para poder predecir si una persona del dataset de test ha sobrevivido o no. A partir de los gráficos del apartado anterior hemos visto clara una relación entre la variables Survived con Sex y Pclass, por lo que estas variables seguro entrarán en el modelo.

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

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex, family = binomial(link = logit),
##     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
```

Nos cuestionamos ahora si más variables podrían hacer mejorar el modelo, como por ejemplo el puerto de embarque

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

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Embarked, family = binomial(link = logit),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3507  -0.6610  -0.4271   0.7227   2.2090
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.6366     0.3322  10.946   <2e-16 ***
## Pclass       -0.9389     0.1110  -8.462   <2e-16 ***
## Sexmale      -2.6192     0.1848 -14.169   <2e-16 ***
## EmbarkedQ    -0.1527     0.3627  -0.421   0.6737
## EmbarkedS    -0.5496     0.2239  -2.454   0.0141 *
## ---
## 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:  820.19  on 886  degrees of freedom
## AIC: 830.19
##
## Number of Fisher Scoring iterations: 4
```

vemos que el factor AIC (Akaike information criterion) ha disminuido, por lo que el modelo mejora. Además los coeficientes del resto de variables no se han visto afectados, lo cual podria indicar una correlación significativa entre las variables explicativas.

Seguimos pues probando con más variables, en este caso el de la edad:

```
M2 <- glm( formula = Survived ~ Pclass + Sex + Embarked + Age, data = train, family=binomial(link=logit
summary(M2)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Embarked + Age, family = binomial(link = logit),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5844  -0.6282  -0.4088   0.6662   2.4758
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.048621   0.481938  10.476  < 2e-16 ***
## Pclass      -1.205798   0.130968  -9.207  < 2e-16 ***
## Sexmale     -2.530510   0.187362 -13.506  < 2e-16 ***
## EmbarkedQ    0.005264   0.371588   0.014   0.9887
## EmbarkedS   -0.475338   0.228420  -2.081   0.0374 *
## Age         -0.032051   0.007388  -4.338 1.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  800.5  on 885  degrees of freedom
## AIC: 812.5
##
## Number of Fisher Scoring iterations: 5
```

vemos que el modelo vuelve a mejorar, disminuyendo el AIC. En este caso vemos que el coeficiente de Pclass ha decrecido, mostrando una relación entre clase y edad que ya vimos previamente. El cambio en el coeficiente no es suficientemente grande como para descartar la variable.

Finalmente nos planteamos incluir también las variables SibSp y/o Parch. Hay una clara relación entre ambas variables por lo que incluir ambas quizás no es buena idea, veamos el AIC que nos dan cada una de ellas:

```
M3 <- glm( formula = Survived ~ Pclass + Sex + Embarked + Age + Parch , data = train, family=binomial(li
M3$aic
```

```
## [1] 811.0777
```

```
M4 <- glm( formula = Survived ~ Pclass + Sex + Embarked + Age + SibSp , data = train, family=binomial(li
M4$aic
```

```
## [1] 798.5619
```

```
M5 <- glm( formula = Survived ~ Pclass + Sex + Embarked + Age + Parch + SibSp, data = train, family=bino
M5$aic
```

```
## [1] 800.2042
```

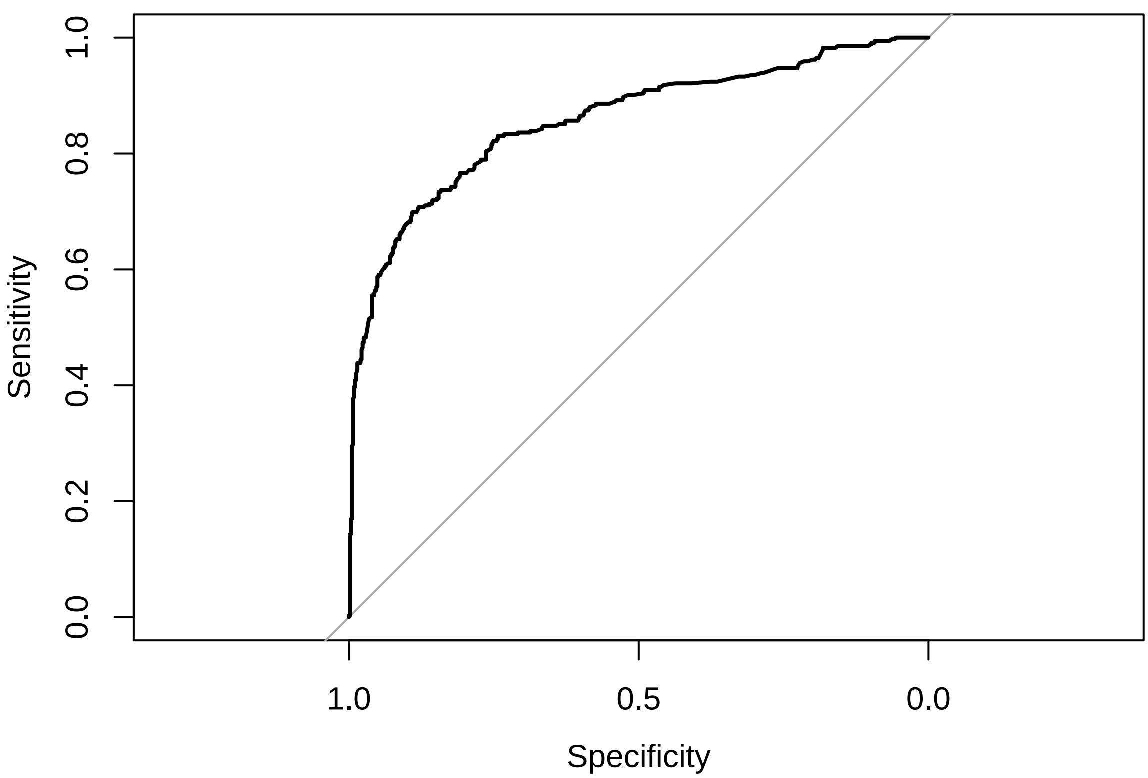
Efectivamente vemos que es mejor quedarse tan solo con la variable SibSp en este caso, por lo que descartamos la variable Parch.

Para terminar, vamos a comprobar lo bien que funciona nuestra regresión con este mismo dataset (más adelante lo veremos con el de test) utilizando la curva ROC

```
prob <- predict(M5, train, type="response")

r <- roc(train$Survived,prob,data=train)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot (r)
```



Esta curva, contra más cerca está de la esquia superior izquierda está, menor es el error que está cometiendo. De hecho podemos extraer un valor númerico que nos dirá con mejor precisión lo bueno que es el modelo:

```
auc(r)
```

```
## Area under the curve: 0.8574
```

Vemos que tenemos un AUC de 0.86. Generalmente se dice que entre 0.6 y 0.8 el modelo se comporta de manera aceptable, pero por encima de 0.8 el modelo se ajusta bien a los datos que intenta reproducir.

Resultados y Conclusiones

Del estudio inicial resulta que las variables Sex, Pclass y Age son las que tienen mayor relación con Survived, puesto que hemos visto en las gráficas de supervivencia vs cada grupo cómo habia una clara relación entre estas variables.

Hay variables que hemos desechado como el número de cabina o el nombre del pasajero puesto que hemos asumido que no tienen relación con la probabilidad de sobrevivir al accidente. Además hemos visto que hay una fuerte relación entre la clase del billete y el precio de éste, por lo que a la hora de proponer la regresión logística hemos decidido descartarla para el modelo.

Haciendo una comparación entre los diferentes modelos de regresión logistica que hemos planteado hemos encontrado que el modelo que mejores resultados arrojaba era el que utilizaba las variables Pclass, Sex, Embarked, Age y SibSp.

Ahora que tenemos ya el modelo construido podemos finalmente participar en la competición propuesta en Kaggle, por lo que vamos a aplicar nuestro modelo al dataset de test y extraer los valores predecidos para la variable Survived:

```
result <- predict(M5, test, type="response")
res_df <- data.frame(test$PassengerId,result)
#Si el resultado es menor a 0.5 le ponemos 0, sino 1
res_df$result_n <- ifelse(res_df$result < 0.5, 0,1)
res_df <- res_df %>% rename( PassengerId = test.PassengerId )
```

Para acabar pasamos a comparar nuestra predicción con los valores reales que se nos aportan en Kaggle para ver si nuestro modelo podría ser utilizado en la competición presentada

```
real_res <- read.table(file="gender_submission.csv",sep=',',dec='.',stringsAsFactors = TRUE,header=TRUE)
pred_vs_real <- merge(real_res,res_df, by=c("PassengerId"), all.x=TRUE)
```

La predicción de nuestro modelo ha acertado un 93.06% de los casos, lo cual consideramos es un resultado muy satisfactorio.