Práctica 2: Limpieza y validación de los datos

Tania Piñeiro y Jordi Sánchez

09/05/2021

Contents

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

El dataset seleccionado es el dataset "Titanic" de Kaggle. Ese conjunto de datos contiene datos sobre los pasajeros de este transantlántico como su edad, género, la clase en la que viajaba y también incluye información sobre si sobrevivió al fatal accidente.

Como es bien conocido, el 15 de abril de 1912, el transatlántico de pasajeros más grande jamás construido chocó con un iceberg durante su viaje inaugural. Cuando el Titanic se hundió, mató a 1502 de los 2224 pasajeros y tripulación. Una de las razones por las que el naufragio resultó en tal pérdida de vidas fue que no había suficientes botes salvavidas para los pasajeros y la tripulación. Aunque hubo algún elemento de

suerte involucrado en sobrevivir al hundimiento, algunos grupos de personas tenían más probabilidades de sobrevivir que otros.

El análisis de este dataset es importante porque puede ofrecer información sobre si hubo diferencias en la supervivencia de los pasajeros en función de sus características como por ejemplo ¿Hubo una mayor probabilidad de fallecidos entre los pasajeros de las clases más bajas? Los resultados nos permitirán obtener conclusiones valiosas sobre el incidente.

Precisamente éste es el problema que pretende resolver el análisis de este dataset, obtener respuestas sobre las características de los pasajeros con mayores posibilidades de sobrevivir.

En primer lugar, vamos a cargar el dataset y analizar de cuantas variables y registros disponemos para abordar este análisis. Se va a cargar el conjunto "train" de entrenamiento disponible en Kaggle.

```
# Cargamos el juego de datos
ds<- read.csv(file='train_titanic.csv',header=T,dec='.', sep=",")
# Verificamos la estructura del conjunto de datos
str(ds)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Se observa que se trata de una base de datos con 891 observaciones y 12 variables en las que se recogen algunas características de los pasajeros. Entre las variables hay 7 variables numéricas y 5 categóricas. A continuación se describen las variables:

- PassengerId (tipo int): código de identificación del pasajero
- Survived (tipo int): informa si el pasajero murió o sobrevivió en el accidente(0 = No, 1 = Si)
- Pclass (tipo int): hace referencia a la clase en la que viajaban los pasajeros (1 = primera clase, 2 = segunda clase...)
- Name (tipo char): nombre del pasajero
- Sex (tipo char): sexo del pasajero
- Age (tipo int): edad del pasajero
- SibSp (tipo int): número de familiares (hermanos o esposa) a bordo del Titanic
- Parch (tipo int): número de familiares (padres o hijos) a bordo del Titanic
- Ticket (tipo char): código del ticket del pasajero
- Fare (tipo num): precio del ticket para viajar en el Titanic
- Cabin (tipo char): número del camarote
- Embarked (tipo char): puerto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton)

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

En el apartado inicial se ha realizado un análisis preliminar de los datos de los que disponemos en el dataset, obteniendo que tiene un total de 12 columnas y 891 registros. Sin embargo, llegado este punto debemos

determinar si realmente necesitamos para nuestro análisis todas esas variables.

Se ha considerado que hay algunas variables que no van a aportar demasiada información al modelo como son: el código del ticket ("Ticket) y el puerto de embarque ("Embarked"). Sin embargo la variable"Ticket" nos permite conocer cuantas personas han comprado el billete con el mismo ticket y de esta forma, calcular el precio por persona de la variable "Fare", que estará agrupada. La variable "Embarked" se va a excluir del análisis. Además, se ha observado que el número de camarote ("Cabin") está ausente en un gran número de registros, sin embargo, por el momento se va a conservar para analizar si se pueden obtener algunas relaciones entre la ubicación del camarote y el desenlace de los pasajeros. El nombre de los pasajeros ("Name"), a priori puede parecer poco relevante pero como incluye el título del pasajero ("Mr., Miss...") se van a conservar por si pudiese ser de utilidad.

De esta forma el dataset que vamos a pre-procesar y analizar contará con 11 columnas y 891 registros.

```
# Selección de las variables de interés
ds<-select(ds, "PassengerId","Name","Survived","Pclass", "Sex","Age","SibSp", "Parch", "Fare", "Cabin",
```

Además, para facilitar el trabajo con estos datos vamos a convertir a factores las variables "Pclass" y "Sex" empleando la función factor. Esto nos facilitará encontrar valores erróneos en futuros pasos y el tratamiento de estas variables para las representaciones gráficas.

```
#La función factor convierte a factor las variables seleccionadas
ds$Pclass<-factor(ds$Pclass)
ds$Sex<-factor(ds$Sex)
ds$Parch<-factor(ds$Parch)
ds$SibSp<-factor(ds$SibSp)
```

Finalmente, se van a obtener las primeras estadísticas descriptivas del conjunto de datos para empezar a conocer más en profundidad como era la distribución de los pasajeros.

```
summary(ds)
```

```
##   PassengerId        Name              Survived      Pclass      Sex
##  Min.   :  1.0   Length:891         Min.   :0.0000   1:216   female:314
##  1st Qu.:223.5   Class :character   1st Qu.:0.0000   2:184   male  :577
##  Median :446.0   Mode  :character   Median :0.0000   3:491
##  Mean   :446.0                      Mean   :0.3838
##  3rd Qu.:668.5                      3rd Qu.:1.0000
##  Max.   :891.0                      Max.   :1.0000
##
##       Age        SibSp   Parch        Fare           Cabin
##  Min.   : 0.42   0:608   0:678   Min.   :  0.00   Length:891
##  1st Qu.:20.12   1:209   1:118   1st Qu.:  7.91   Class :character
##  Median :28.00   2: 28   2: 80   Median : 14.45   Mode  :character
##  Mean   :29.70   3: 16   3:  5   Mean   : 32.20
##  3rd Qu.:38.00   4: 18   4:  4   3rd Qu.: 31.00
##  Max.   :80.00   5:  5   5:  5   Max.   :512.33
##  NA's   :177     8:  7   6:  1
##     Ticket
##  Length:891
##  Class :character
##  Mode  :character
##
##
```

```
##
##
```

De los resultados obtenidos se obtienen las siguientes conclusiones:

- De los 891 pasajeros, 549 no sobrevivieron al accidente y 342 sí lo hicieron.
- De los 891 pasajeros, 216 iban en primera clase, 184 en segunda y 491 en tercera.
- De los 891 pasajeros, 314 eran mujeres y 577 hombres.
- La edad de los pasajeros se encuentra entre los 0.42 (posible error) y los 80 años, siendo la edad media 29.7
- Los pasajeros tenían de 0-8 hermanos/esposas a bordo y entre 0-6 hijos/padres.

3. Limpieza de los datos

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

El siguiente paso consistirá en determinar si hay valores nulos y si es así eliminarlos o sustituirlos por otros.

```
# Estadísticas de valores vacíos
colSums(is.na(ds)) %>% kable(caption="Número de NAs por columna") %>% kable_styling(latex_options = "ho
```

Table 1: Número de NAs por columna

	x
PassengerId	0
Name	0
Survived	0
Pclass	0
Sex	0
Age	177
SibSp	0
Parch	0
Fare	0
Cabin	0
Ticket	0

```
colSums(ds=="", na.rm = TRUE) %>% kable(caption="Número de valores vacíos por columna") %>% kable_stylin
```

En las tablas superiores se han obtenido el número de registros con valor "NA" para cada columna y también el número de valores vacíos. En la primera tabla se observa que la única columna se muestra que la única variable que presenta valores ausentes es "Age" con 177 registros NA.

En la segunda tabla se observa que la única variable que tiene valores vacíos es "Cabin" con 687 registros vacíos. Este es un número muy elevado teniendo en cuenta que nuestro dataset cuenta con 897 registros, significa que el 77% de los registros de esta columna no están disponibles.

En cuanto a la gestión que se va a hacer de estos valores faltantes hemos determinado conservar la variable "Age" ya que consideramos que puede ser muy relevante para el análisis. Se va a realizar una imputación de los valores faltantes, obteniendo la edad media para cada uno de los grupos de títulos de los pasajeros

Table 2: Número de valores vacíos por columna

	x
PassengerId	0
Name	0
Survived	0
Pclass	0
Sex	0
Age	0
SibSp	0
Parch	0
Fare	0
Cabin	687
Ticket	0

("Mr, Miss"). Es decir, para una edad faltante de una pasajera tipo "Miss" se le inputará la edad promedio de todas las pasajeras tipo "Miss" del dataset.

Por otra parte, se decide finalmente eliminar la variable "Cabin" ya que se considera que un 77% de datos faltantes es demasiado elevado como para intentar algún tipo de imputación y se podría introducir error.

```
# Se elimina la variable "Cabin"
ds <- ds[,-(10)]
```

Como se ha comentado la edad se va a estimar en función del título del pasajero. Para ello es necesario, previamente separar el título de la variable "Name" y recogerlo en una variable independiente "Title". Para realizar esta separación se va a emplear una expresión regular, aprovechando que el título aparece tras una coma y un espacio y antes de un punto (Ej: "Braund, Mr. Owen Harris"). A continuación se va a identificar cuántos tipos de títulos hay y su recuento.

```
# Separación del título del pasajero en una nueva columna
ds$Title <- gsub('(.*, )|(\\..*)', '', ds$Name)
table(ds$Title)
```

```
##
##         Capt          Col          Don           Dr     Jonkheer         Lady
##            1            2            1            7            1            1
##        Major       Master         Miss         Mlle          Mme           Mr
##            2           40          182            2            1          517
##          Mrs           Ms          Rev          Sir the Countess
##          125            1            6            1            1
```

Se obtiene que hay un total de 17 títulos diferente, aunque la mayoría de ellos son poco frecuentes y solo se encuentran en un par de registros como "Capt", "Major". Estos títulos poco frecuentes se van a agrupar en un único factor, de forma que finalmente habrá 5 niveles: "Miss", "Master", "Mrs", "Mr" y "Other".

```
# Se agrupan los títulos poco frecuentes
other <- c('Dona', 'Lady', 'the Countess','Capt', 'Col', 'Don',
                'Dr', 'Major', 'Rev', 'Sir', 'Jonkheer', 'Mlle', 'Ms', 'Mme', 'Lady')
ds$Title[ds$Title %in% other]  <- 'Other'
# Se eliminan los niveles no empleados
ds$Title<-factor(ds$Title)
table(ds$Title)
```

```
##
## Master   Miss     Mr    Mrs  Other
##     40    182    517    125     27
```

Ahora observamos que hemos obtenido los cinco grupos de títulos de los pasajeros. Vamos a aprovechar esta información para realizar la inputación de los valores faltantes como el valor promedio de estos grupos.

```
# Se localizan los na de las variables Weight y Height
i_na<-is.na(ds$Age)
dcomplete<-ds[!i_na,]
# Se estima el promedio de edad para cada uno de los grupos
av_Age_title<-aggregate(x = dcomplete$Age, by = list(dcomplete$Title), FUN = mean)
# Imputación
ds[which(is.na(ds$Age)&ds$Title=="Master"),"Age"] <- round(av_Age_title[1,2])
ds[which(is.na(ds$Age)&ds$Title=="Miss"),"Age"] <- av_Age_title[2,2]
ds[which(is.na(ds$Age)&ds$Title=="Mr"),"Age"] <- av_Age_title[3,2]
ds[which(is.na(ds$Age)&ds$Title=="Mrs"),"Age"] <- av_Age_title[4,2]
ds[which(is.na(ds$Age)&ds$Title=="Other"),"Age"] <- av_Age_title[5,2]
colSums(is.na(ds))
```

```
## PassengerId        Name    Survived      Pclass         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch        Fare      Ticket       Title
##           0           0           0           0           0
```

Comprobamos que tras realizar la imputación de valores a la variable "Age" ya no hay ningun registro con valores NA. De esta forma confirmamos que se ha realizado adecuadamente la imputación de valores. Se podrían haber realizado otros procedimientos para el tratamiento de estos datos faltantes como la eliminación de toda la fila donde se encuentre un registro faltante o la imputación directamente de la edad promedio global. Sin embargo, el número de valores faltantes (177) se consideró demasiado elevado como para eliminar los registros, ya que se perdería gran cantidad de información. Además, la variable age tiene un rango bastante amplio (0.4 - 80 años) de forma que si imputásemos el promedio global de todos los pasajeros probablemente el error que estaríamos introduciendo sería mayor.

3.2. Identificación y tratamiento de valores extremos.

Para el tratamiento de valores extremos trabajaremos con las variables "Age", "Fare", "SibSp" y "Parch", ya que son las únicas variables númericas de las que disponemos en el dataset.

Sin embargo, para empezar, verificaremos que entre las variables "Survived", "Pclass", "Sex", "SibSp" y "Parch" que son tipo factor no hay ningún nivel que pueda ser anómalo.

```
# Comprobación de las variables tipo factor
summary(ds[c("Survived", "Pclass", "Sex", "SibSp","Parch")])
```

```
##     Survived      Pclass      Sex      SibSp   Parch
##  Min.   :0.0000   1:216   female:314   0:608   0:678
##  1st Qu.:0.0000   2:184   male  :577   1:209   1:118
##  Median :0.0000   3:491                2: 28   2: 80
##  Mean   :0.3838                        3: 16   3:  5
##  3rd Qu.:1.0000                        4: 18   4:  4
##  Max.   :1.0000                        5:  5   5:  5
##                                        8:  7   6:  1
```

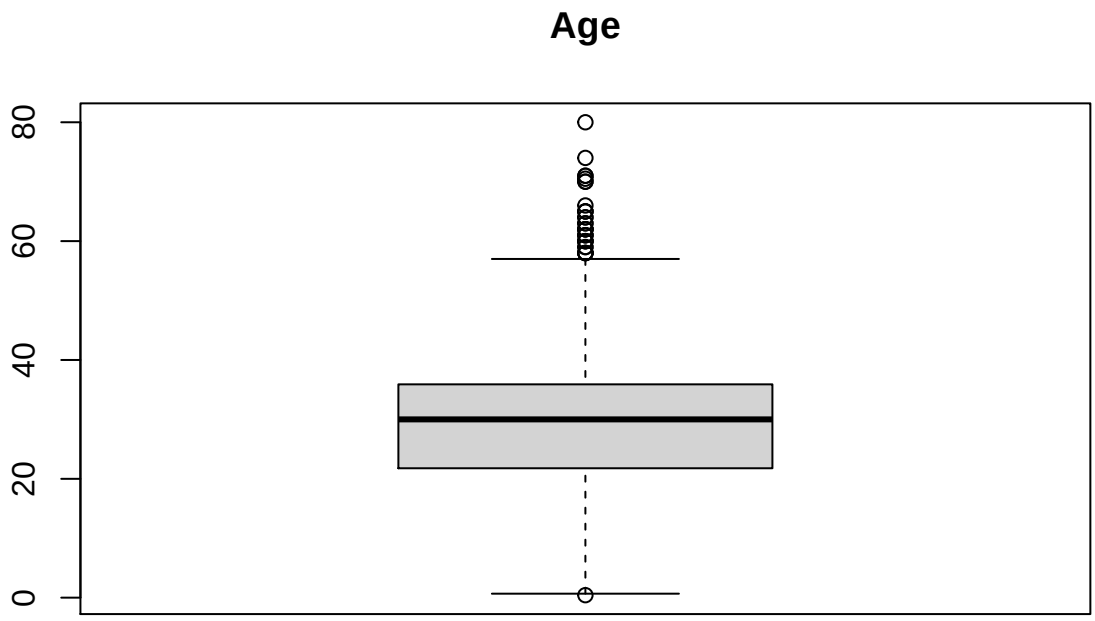
En los resultados obtenidos se comprueba que no hay ningún valor extremo en estas variables:

Survived: se comprueba que toma valores 0 o 1. **Pclass:** toma valores entre 1 y 3. **Sex:** se comprueba que hay dos niveles "female" y "male" **SibSp:** toma valores entre 0 - 8, siendo los más frecuentes 0 (608 pasajeros) y 1 (209 pasajeros) **Parch:** toma valores entre 0 - 6, siendo los más frecuentes 0 (678 pasajeros) y 1 (118 pasajeros)

Vamos ahora a verificar si en las variables "Age" y "Fare" hay valores extremos. Si es así, y se trata de un valor anormalmente alto o bajo, se sustituirá el valor por "NA", para realizar posteriormente una posible imputación. Para localizar estos valores extremos se va a emplear la representación boxplot que permite obtener los outliers (valores atípicos) de una determinada variable. Empleando "$out" obtenemos estos valores extremos que son posteriormente sustituidos por NA en las variables originales.

Age

```
# Se representan el boxplot de la variable Age
b1<-boxplot(ds$Age, main="Age")
```




```
# Se obtienen las estadísticas
b1$stats
```

```
##          [,1]
## [1,]  0.67000
## [2,] 21.77397
## [3,] 30.00000
## [4,] 35.89815
## [5,] 57.00000
```

```r
# Se contabilizan los outliers
length(b1$out)
```

```
## [1] 34
```

```r
summary(ds[,"Age"])
```

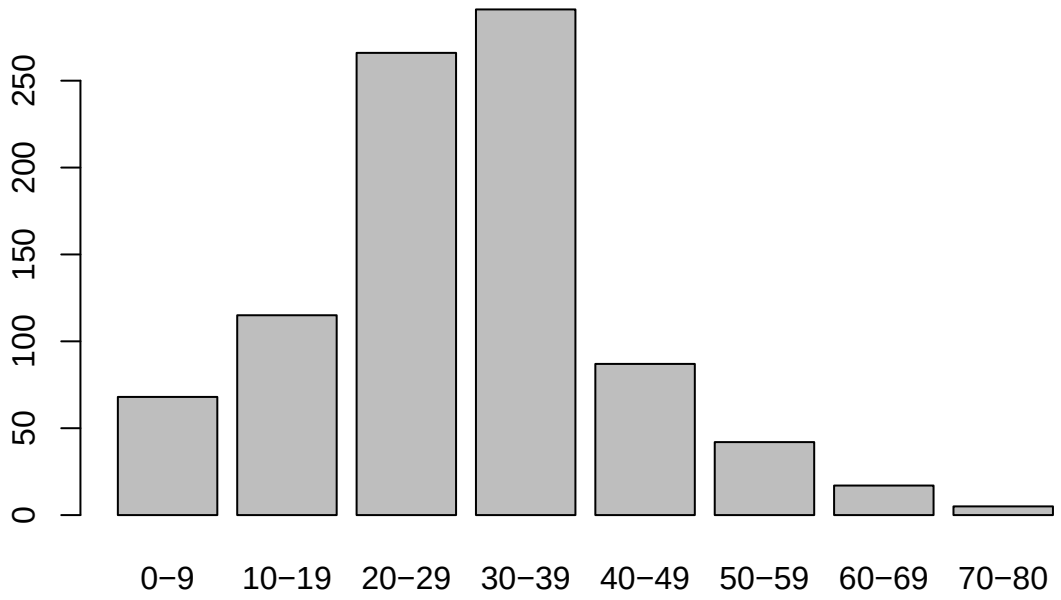
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   21.77   30.00   29.76   35.90   80.00
```

Para la variable Age se obtiene que el valor del bigote inferior es 0.67 y el del bigote superior 57.0. Los extremos de la caja son 21.77 y 35.89, inferior y superior respectivamente, siendo la mediana de la edad de los pasajeros 30.0 años. La interpretación de estos resultados indica que la mitad de los pasajeros a bordo del titanic tenía entre 21.77 y 35.89 años. Además se ha obtenido que hay 34 puntos considerados outliers por situarse alejados del resto de datos, 1.5 veces menor o mayor que los extremos de los bigotes. Sin embargo si observamos el rango de la edad de los pasajeros observamos que los valores se encuentran acotados entre 0.42 y 80 años, que son edades razonables. De forma que no se van a eliminar estos outliers, porque forman parte de la diversidad de la muestra. Podrían aplicarse otros procedimientos como la eliminación de las filas con outliers o realizar imputación de estos valores como hicimos en el punto anterior. En este caso se va a realizar la discretización de la variable en grupos de edad, para reducir el ruido de la misma.

```r
# Discretizamos
ds["segmento_edad"] <- cut(ds$Age, breaks = c(0,10,20,30,40,50,60,70,100), labels = c("0-9", "10-19", "
# Observamos los datos discretizados.
head(ds)
```

```
##   PassengerId                                                 Name Survived
## 1           1                              Braund, Mr. Owen Harris        0
## 2           2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)        1
## 3           3                               Heikkinen, Miss. Laina        1
## 4           4       Futrelle, Mrs. Jacques Heath (Lily May Peel)        1
## 5           5                             Allen, Mr. William Henry        0
## 6           6                                     Moran, Mr. James        0
##   Pclass    Sex      Age SibSp Parch    Fare           Ticket Title
## 1      3   male 22.00000     1     0  7.2500        A/5 21171    Mr
## 2      1 female 38.00000     1     0 71.2833         PC 17599   Mrs
## 3      3 female 26.00000     0     0  7.9250 STON/O2. 3101282  Miss
## 4      1 female 35.00000     1     0 53.1000           113803   Mrs
## 5      3   male 35.00000     0     0  8.0500           373450    Mr
## 6      3   male 32.36809     0     0  8.4583           330877    Mr
##   segmento_edad
## 1         20-29
## 2         30-39
## 3         20-29
## 4         30-39
## 5         30-39
## 6         30-39
```

```r
# Vemos como se agrupan los datos.
plot(ds["segmento_edad"])
```



Se ha discretizado la variable "Age" en ocho intervalos de 10 años cada uno. Esta segmentación será de utilidad en apartados posteriores para analizar las relaciones de los datos. En el gráfico de barras representado se observa que, como ya habíamos obtenido la mayoría de los pasajeros se concentran en los segmentos "20-29" y "30-39".

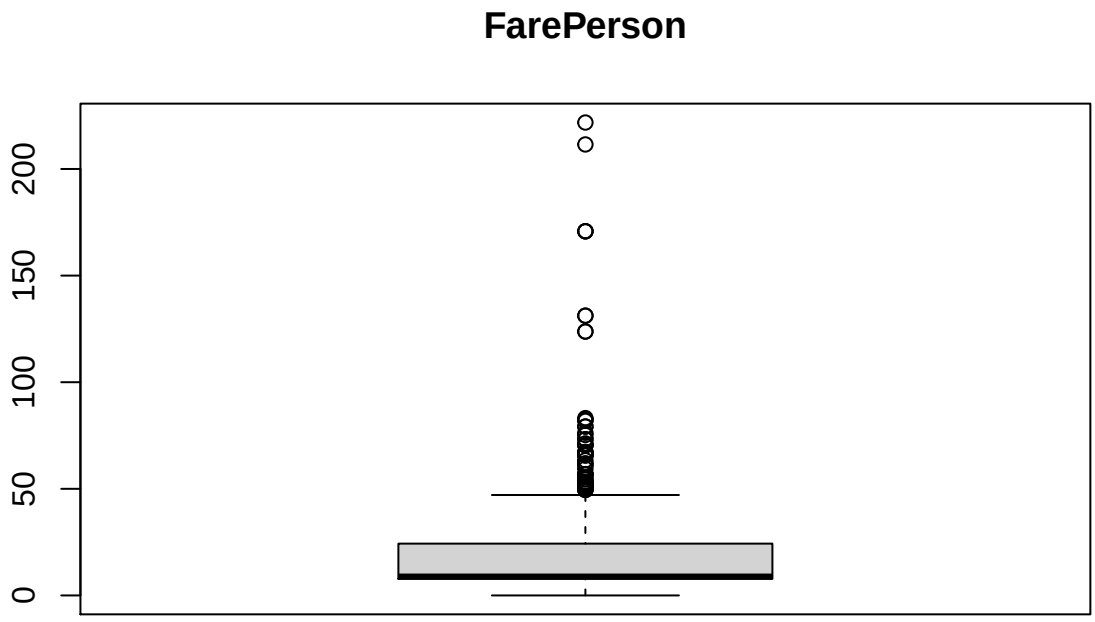
Fare

Se va a realizar el mismo análisis para la variable Fare para tratar de determinar si hay valores extremos en esta variable. En primer lugar, aunque el atributo Ticket es algo inútil en cuanto a extraer información del propio número de ticket, proporciona información sobre cuántos tickets se compraron con una tarifa determinada, de modo que podamos calcular la tarifa por persona, que es lo que necesitamos ya que nuestra unidad de observación (una fila) es una persona. Creamos un atributo FarePerson y lo comparamos con el atributo Fare:

```
# Cuenta de tickets
counts <- aggregate(ds$Ticket, by=list(ds$Ticket),
                    FUN=function(ticket) sum(!is.na(ticket)))
# Función para el cálculo de ratio de tickets
compute_fare_person <- function(ds) {
  fare <- as.numeric(ds["Fare"])
  # Cuenta
  count_ticket_i <- counts[which(counts[,1] == ds["Ticket"]), 2]
  result <- round(fare/count_ticket_i,2)
  return(result)
}
# Aplicamos la función para obtener FarePerson
ds$FarePerson <- apply(X=ds, MARGIN=1, FUN=compute_fare_person)
```

A continuación, se representará el diagrama de cajas y bigotes y finalmente se analizarán los quartiles obtenidos y los posibles outliers.

```
# Se representan los boxplot de la variable Fare
b2<-boxplot(ds$FarePerson, main="FarePerson")
```




```
b2$stats
```

```
##        [,1]
## [1,]  0.000
## [2,]  7.765
## [3,]  8.850
## [4,] 24.290
## [5,] 47.100
```

```
length(b2$out)
```

```
## [1] 59
```

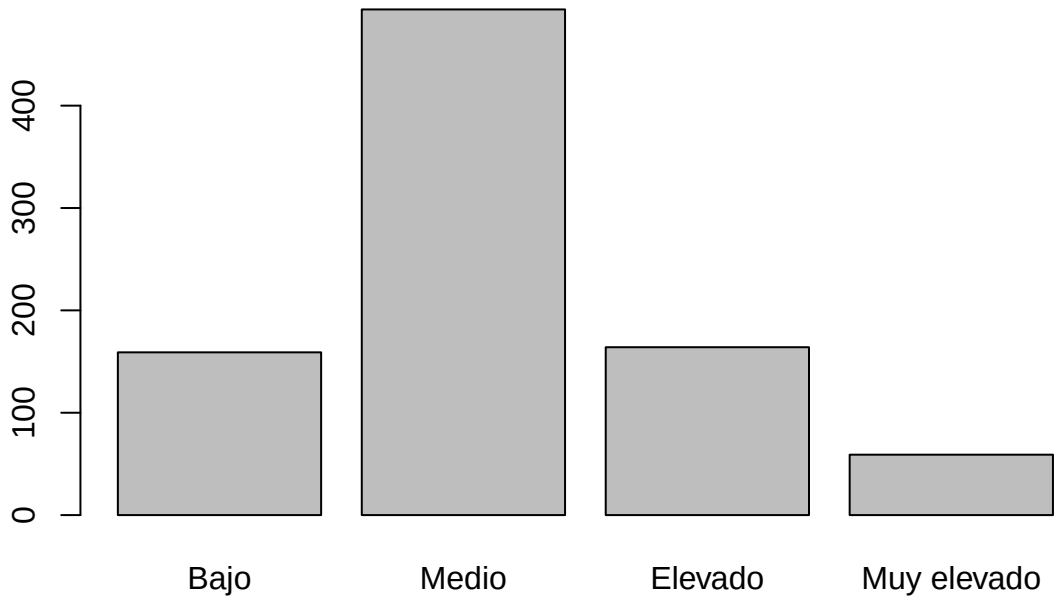
Para la variable Age se obtiene que el valor del bigote inferior es 0 y el del bigote superior 47.10 dólares. Los extremos de la caja son 7.765 y 24.29 dólares, inferior y superior respectivamente, siendo la mediana del precio del ticket por pasajero 8.85 dólares. La interpretación de estos resultados indica que la mitad de los pasajeros a bordo del titanic pagó entre 7.765 y 24.29 dólares por el billete. Además se ha obtenido que hay 59 puntos considerados outliers por situarse alejados del resto de datos, 1.5 veces mayor que el extremo de

los bigotes. Para tratar estos valores extremos podrían aplicarse procedimientos como la eliminación de las filas con outliers o realizar imputación de estos valores como hicimos en el punto anterior. En este caso se va a realizar la discretización de la variable en grupos de precio, para reducir el ruido de la misma.

```
# Discretizamos
ds["segmento_fare"] <- cut(ds$FarePerson, breaks = c(0,7.7,24.3,47.10,300), labels = c("Bajo", "Medio",
# Observamos los datos discretizados.
head(ds)
```

```
##   PassengerId                                                Name Survived
## 1           1                             Braund, Mr. Owen Harris        0
## 2           2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)        1
## 3           3                              Heikkinen, Miss. Laina        1
## 4           4        Futrelle, Mrs. Jacques Heath (Lily May Peel)        1
## 5           5                            Allen, Mr. William Henry        0
## 6           6                                    Moran, Mr. James        0
##   Pclass    Sex      Age SibSp Parch    Fare           Ticket Title
## 1      3   male 22.00000     1     0  7.2500        A/5 21171    Mr
## 2      1 female 38.00000     1     0 71.2833         PC 17599   Mrs
## 3      3 female 26.00000     0     0  7.9250 STON/O2. 3101282  Miss
## 4      1 female 35.00000     1     0 53.1000           113803   Mrs
## 5      3   male 35.00000     0     0  8.0500           373450    Mr
## 6      3   male 32.36809     0     0  8.4583           330877    Mr
##   segmento_edad FarePerson segmento_fare
## 1         20-29       7.25          Bajo
## 2         30-39      71.28   Muy elevado
## 3         20-29       7.92         Medio
## 4         30-39      26.55       Elevado
## 5         30-39       8.05         Medio
## 6         30-39       8.46         Medio
```

```
# Vemos como se agrupan los datos.
plot(ds["segmento_fare"])
```



Se ha discretizado la variable "FarePerson" en cuatro intervalos (Bajo, Medio, Elevado y Muy elevado) en función de los rangos intercuartílicos obtenidos en el gráfico de caja y bigotes. . Esta segmentación será de utilidad en apartados posteriores para analizar las relaciones de los datos. En el gráfico de barras representado se observa la distribución del número de billetes de cada categoría.

4. Análisis de los datos.

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

En este apartado se va a realizar un diseño del estudio estadístico que se quiere desarrollar. El objetivo último es definir si hay diferencias en la variable de clasificación "Survived", es decir en el resultado de supervivencia del pasajero, entre los diferentes grupos de pasajeros.

De las variables disponibles en el dataset se van a incluir en el estudio: La variable **"Sex"**, categórica, para determinar si hay diferencias entre hombres y mujeres en la supervivencia. En el conjunto de entrenamiento hay un total de 314 mujeres y 577 hombres.

```
table(ds$Sex)
```

```
##
## female   male
##    314    577
```

La variable **"Age_segment"**, para determinar si hay diferencias entre los distintos rangos de edad en la supervivencia. En la tabla inferior se muestra el número de pasajeros que hay en cada grupo de edad, se observa que los grupos más numerosos son los correspondientes a los intervales "20-29" y "30-39" años.

```
table(ds["segmento_edad"])
```

```
##
##   0-9 10-19 20-29 30-39 40-49 50-59 60-69 70-80
##    68   115   266   291    87    42    17     5
```

La variable **FarePerson** para tratar de determinar si aquellos pasajeros que pagaron un ticket más caro tenían más posibilidades de supervivencia o a la inversa.La gran mayoría de los pasajeros compraron un billete con un precio medio.

```
table(ds["segmento_fare"])
```

```
##
##        Bajo       Medio     Elevado Muy elevado
##         159         494         164          59
```

La variable **Pclass** para determinar si hay diferentes en la superviviencia entre los pasajeros que viajaban en diferentes clases. Como se observa en la tabla inferior, el grupo más numeroso es el de pasajeros que viajaban en tercera clase.

```
table(ds$Pclass)
```

```
##
##   1   2   3
## 216 184 491
```

Finalmente se trabajará con la variable **nfamilares**, que representará el número de familiares que tenía cada pasajero a bordo del Titanic. Esta variable se obtendrá como suma de la variable "SibSp" y "Parch", de forma que no se diferenciará por el tipo de familiar si no por el número total de familiares. De esta forma se tratará de determinar si hay una mayor supervivencia en los pasajeros que tenían más (o menos) familiares a bordo.Como se observa en la tabla inferior, la mayoría de los pasajeros tenían 2 familiares a bordo, seguido de 3 y 4.

```
ds$nfamiliares=as.numeric(ds$SibSp)+as.numeric(ds$Parch)
table(ds$nfamiliares)
```

```
##
##   2   3   4   5   6   7   8   9  10
## 537 161 102  29  15  22  12   6   7
```

De esta forma, se concluye que las variables que se van a analizar en los siguientes apartados para estudiar su correlación con la supervivencia de los pasajeros son: "Sex", "Age_segment", "Fare_person","Pclass" y "nfamiliares".

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

Primero lo que haremos será agrupar estas variables a analizar en un nuevo dataset y aplicar One Hot Encoding en las variables categóricas. Para la comprobación de la normalidad y homogeneidad utilizaremos la prueba de **Anderson-Darling** con tal de verificar que la variable en cuestión siga una distribución normal.

```
library(mltools)
```

```
## Warning: package 'mltools' was built under R version 4.0.5
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.0.3
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
incluir <- c("Survived","Sex","segmento_edad","FarePerson","Pclass","nfamiliares")
newds <- ds[ , (names(ds) %in% incluir)]

newdata <- one_hot(as.data.table(newds["segmento_edad"]))
newds <- cbind(newds, newdata)
newdata <- one_hot(as.data.table(newds["Sex"]))
newds <- cbind(newds, newdata)
newdata <- one_hot(as.data.table(newds["Pclass"]))
newds <- cbind(newds, newdata)

newds <- newds[ , !(names(newds) %in% c("segmento_edad","Sex","Pclass"))]
```

```
library(nortest)
```

```
## Warning: package 'nortest' was built under R version 4.0.3
```

```
alpha = 0.05
col.names = colnames(newds)
for (i in 1:ncol(newds)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(newds[,i]) | is.numeric(newds[,i])) {
    p_val = ad.test(newds[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
```

```
      cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## Survived
## FarePerson
## nfamiliares
## segmento_edad_0-9
## segmento_edad_10-19
## segmento_edad_20-29
## segmento_edad_30-39
## segmento_edad_40-49
## segmento_edad_50-59
## segmento_edad_60-69
## segmento_edad_70-80
## Sex_female
## Sex_male
## Pclass_1
## Pclass_2
## Pclass_3
```

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

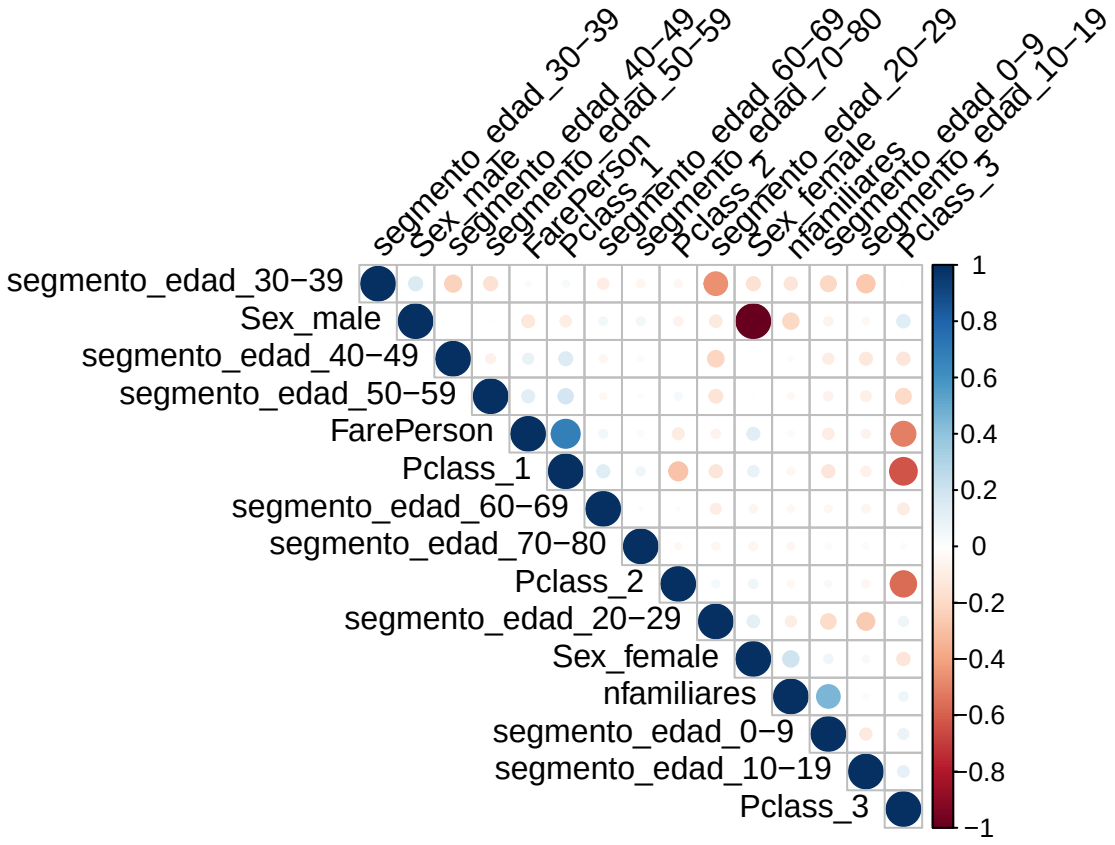
Primero vamos a ver que variables influyen más en el resultado a partir de la correlación de estas.

```
#newds[ , !(names(newds) == "Survived")]
res <- cor(newds[ , !(names(newds) == "Survived")])
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.0.3
```

```
## corrplot 0.84 loaded
```

```
corrplot(res, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Veamos ahora que tal se adapta un modelo de regresión logistica a estos datos.

```
train <- newds[1:667,]
test <- newds[668:891,]

## Model Creation
model <- glm(Survived ~.,family=binomial(link='logit'),data=train)

## Model Summary
summary(model)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6957  -0.6560  -0.4690   0.5776   2.3845
##
## Coefficients: (3 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -2.2366274  1.2340224  -1.812 0.069914 .
## FarePerson            -0.0002408  0.0062941  -0.038 0.969483
## nfamiliares           -0.3150875  0.0887699  -3.549 0.000386 ***
## `segmento_edad_0-9`    2.5153192  1.3051901   1.927 0.053959 .
```

```
## `segmento_edad_10-19`  0.9250455  1.2560714   0.736 0.461451
## `segmento_edad_20-29`  0.7227972  1.2288784   0.588 0.556414
## `segmento_edad_30-39`  0.7169773  1.2217571   0.587 0.557310
## `segmento_edad_40-49`  0.2952921  1.2460357   0.237 0.812668
## `segmento_edad_50-59` -0.0007082  1.2879037  -0.001 0.999561
## `segmento_edad_60-69` -0.4654978  1.4011504  -0.332 0.739719
## `segmento_edad_70-80`         NA         NA      NA       NA
## Sex_female             2.8113770  0.2311331  12.163  < 2e-16 ***
## Sex_male                      NA         NA      NA       NA
## Pclass_1               2.1010875  0.3593706   5.847 5.02e-09 ***
## Pclass_2               1.1488504  0.2642347   4.348 1.37e-05 ***
## Pclass_3                      NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 891.99  on 666  degrees of freedom
## Residual deviance: 600.41  on 654  degrees of freedom
## AIC: 626.41
##
## Number of Fisher Scoring iterations: 5
```

```r
## Using anova() to analyze the table of devaiance
anova(model, 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                                   666     891.99
## FarePerson            1   33.073       665     858.92 8.878e-09 ***
## nfamiliares           1    0.312       664     858.61 0.5762877
## `segmento_edad_0-9`   1   11.637       663     846.97 0.0006467 ***
## `segmento_edad_10-19` 1    1.332       662     845.64 0.2484880
## `segmento_edad_20-29` 1    6.058       661     839.58 0.0138420 *
## `segmento_edad_30-39` 1    1.500       660     838.08 0.2206139
## `segmento_edad_40-49` 1    1.526       659     836.56 0.2167781
## `segmento_edad_50-59` 1    0.833       658     835.72 0.3613342
## `segmento_edad_60-69` 1    0.000       657     835.72 0.9890117
## `segmento_edad_70-80` 0    0.000       657     835.72
## Sex_female            1  194.739       656     640.98 < 2.2e-16 ***
## Sex_male              0    0.000       656     640.98
## Pclass_1              1   21.385       655     619.60 3.758e-06 ***
## Pclass_2              1   19.188       654     600.41 1.184e-05 ***
## Pclass_3              0    0.000       654     600.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Predicting Test Data
result <- predict(model,newdata=test,type='response')
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
result <- ifelse(result > 0.5,1,0)

result<-factor(result)
test$Survived<-factor(test$Survived)


## Confusion matrix and statistics
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.0.3
```

```
## Loading required package: lattice
```

```r
confusionMatrix(data=result, reference=test$Survived)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 130  28
##          1  12  54
##
##                Accuracy : 0.8214
##                  95% CI : (0.7649, 0.8693)
##     No Information Rate : 0.6339
##     P-Value [Acc > NIR] : 6.393e-10
##
##                   Kappa : 0.5987
##
##  Mcnemar's Test P-Value : 0.01771
##
##             Sensitivity : 0.9155
##             Specificity : 0.6585
##          Pos Pred Value : 0.8228
##          Neg Pred Value : 0.8182
##              Prevalence : 0.6339
##          Detection Rate : 0.5804
##    Detection Prevalence : 0.7054
##       Balanced Accuracy : 0.7870
##
##        'Positive' Class : 0
##
```

```
## ROC Curve and calculating the area under the curve(AUC)
library(ROCR)
```

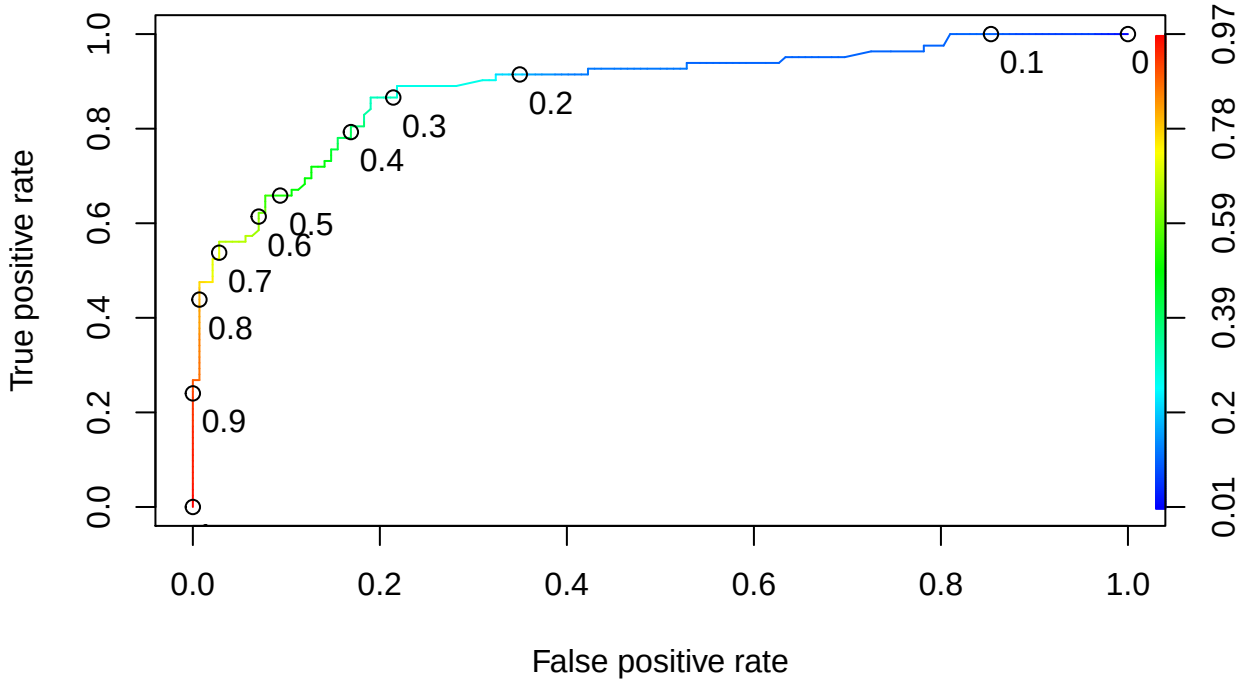
```
## Warning: package 'ROCR' was built under R version 4.0.3
```

```
predictions <- predict(model, newdata=test, type="response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
ROCRpred <- prediction(predictions, test$Survived)
ROCRperf <- performance(ROCRpred, measure = "tpr", x.measure = "fpr")

plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7), print.cutoffs.at = seq(0,1,0.1))
```



5. Representación de los resultados a partir de tablas y gráficas.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?