Práctica 2: Limpieza y análisis de datos

Autor: Adonis González Godoy, Eduardo Tremps Pallarés

Junio 2020

Contents

1 Detalle de la actividad

1.1 Descripción

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2 Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.3 Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2 Resolución

2.1 Descripción del dataset

El conjunto de datos se ha obtenido de análisis se ha obtenido desde **Kaggle**, se puede acceder a través del siguiente link: https://www.kaggle.com/c/titanic, este dataset esta constituido por [891] filas de observaciones y [12] variables (columnas).

A continuación se describen las variables contenidas en el fichero:

Variables	Descripción
passengerId	- int, valor de identificación único de cada pasajero
name	- string, que hace referencia al nombre del pasajero
sex	- factor, con niveles (masculino y femenino)
age	- numeric, valor que se refiere a la edad de una persona determinada. La edad de los niños menores de 12 meses es dad en fracción de un año (1/mes)
class	- factor, especifica la clase para cada pasajero (tipo de servicio a bordo)
embarked	- factor, hace referencia al lugar de embarcamiento (puerto de embarque de las personas)
ticketno	- numeric, especifica el número de ticket (na para la tripulación)
fare	- numeric, valor con el precio del ticket (na para la tripulación, musicos, empleados y otros)
sibsp	- factor ordenado, especifíca el número de hermanos/familiares
cabin	- factor, tipo de cabina que ocupa cada pasajero
parch	- factor ordenado, especifica el número de padres e hijos a bordo
survived	- factor 2 de dos niveles, que especifíca (sí o no) la persona ha sobrevivido al hundimiento

¿Por qué es importante y qué problema pretende resolver?

El hundimiento del Titanic es uno de los naufragios más tristes conocidos de la historia, Titanic fue considerado "insumergible" a pesar del impacto con el iceberg, como resultado dejando muertes de pasajeros y tripulación.

Aunque había un elemento que hacía inclinar a favor, a la hora de supervivencia de algunos grupos de personas, tenían más probabilidades de sobrevivir que otros, este elemento o variables son los que se intentarán resolver o llegar alguna conclusión en este caso práctico.

2.2 Integración y selección de los datos

A partir de este conjunto de datos se intentará resolver qué variables o atributos son los que más influyeron a la hora de la supervivencia de los pasajeron y tripulación en el Titanic. Este conjunto, como se mencionó contiene más de 891 filas por 12 columnas, los datos son de tipo variado por lo que hacen un dataset muy interesante para este análisis.

Para la representación de resultados se utilizará difentes tipos de gráficos que ayudarán a obtener conclusiones claras y directas.

2.3 Limpieza de los datos.

Primero de todo se procede a leer el fichero de datos con formato CSV, para esto se utiliza la función read.csv(), además mostramos las primeras cinco filas del dataset para tener un primer contacto.

```
# Cargamos el juego de datos
titanicData <- read.csv('train.csv', header = TRUE)

# Mostramos las 5 primeras filas del dataset
head(titanicData, 5)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
##                                                 Name    Sex Age SibSp Parch
## 1                            Braund, Mr. Owen Harris   male  22     1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0
## 3                             Heikkinen, Miss. Laina female  26     0     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0
## 5                           Allen, Mr. William Henry   male  35     0     0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
```

Antes de proceder con limpieza de datos, es importante conocer con qué tipo de datos estamos tratando, para esto realizamos las estadísticas básicas del conjunto de datos.

Tenemos la dimesión.

```
# Dimesión del dataset
dim(titanicData)
```

```
## [1] 891  12
```

Mostramos las estadísticas básicas.

```
# Estadísticas básicas
summary(titanicData)
```

```
##   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
```

Aquí ya se puede observar que disponemos de datos NA en el campo edad, disponemos de 177 valores NA.

Mostramos la estructura del conjunto de datos.

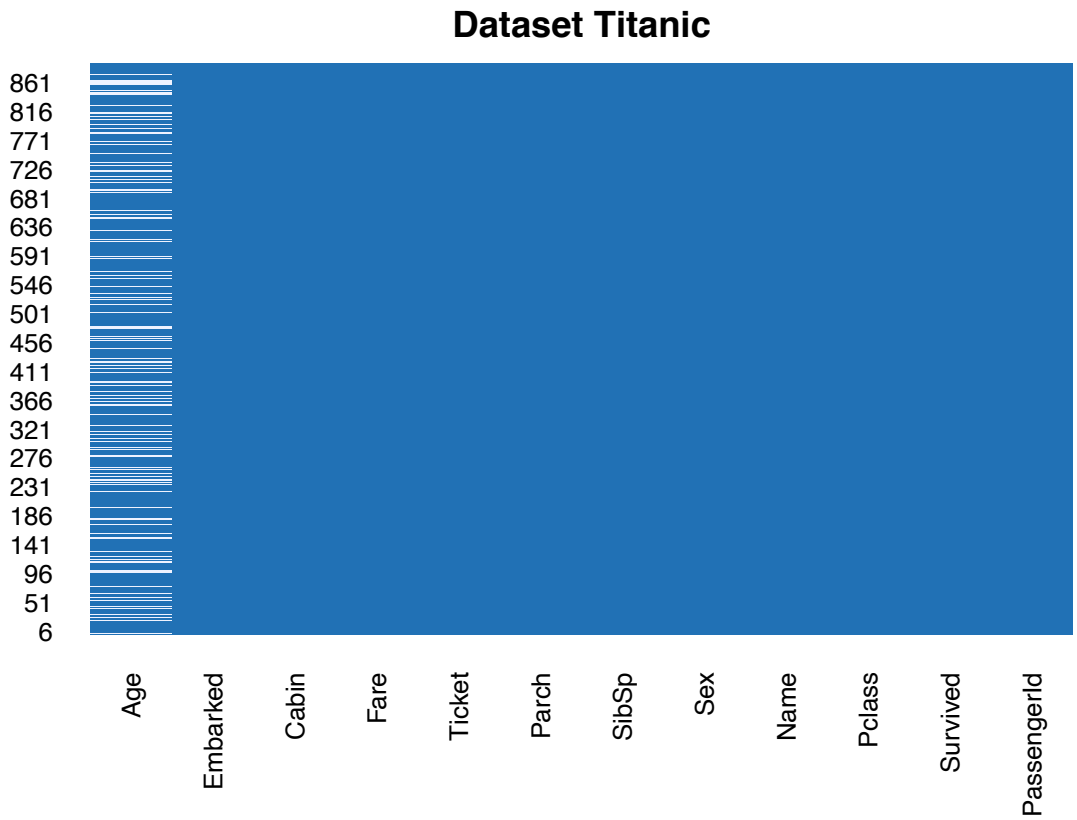
```
# Verificamos la estructura del conjunto de datos
str(titanicData)
```

```
## '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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 58
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

2.3.1 Ceros y elementos vacíos

Primero de todo, comprobaremos si existen valores NA usando la función missmap() de la libería "Amelia", nos permite de una manera visual obtener que variables con datos nulos.

```
# Importamos la libreria amelia
library(Amelia)

# Pintamos el diagrama de valores faltantes
missmap(obj =titanicData, main ="Dataset Titanic", legend =FALSE)
```



Se procede a comprobar la cantidad de elementos nulos, para comprobar los elementos nulos también podemos usar la funcion is.na().

```
# Estadísticas de valores vacíos
colSums(is.na(titanicData))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
```

Podemos observar que la variable edad contiene datos nulos pero aún no hemos revisado los campos de strings vacíos.

Se procede a comprobar si existen valores vacíos, usando la función colSums(), esta función devuelve el número de elementos vacíos.

```
# Estadísticas de valores vacíos
colSums(titanicData=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

```
# Verificamos la cantidad de elementos solo por el atributo cabin
length(which(titanicData$Cabin==""))
```

```
## [1] 687
```

Se puede observar que también disponemos de valores vacíos, el tipo de cabina (cabin) de cada pasajero contiene 684 elementos vacíos y estos no han sido identificados como nulos, en la gráfica anterior se observa que solo teníamos datos nulos en edad.

```
sinCabina <-which(titanicData$Pclass==1 & titanicData$Cabin=="")
length(sinCabina)
```

```
## [1] 40
```

En el dataset tenemos 3 tipos de clases (1,2 y3) y solo a los de primera clase tenían una cabina, por lo que las observaciones obtenidas para los pasajeros de primera clase que no contienen cabina son datos que faltan.

Podremos valores nulos ya que se ha verificado realmente son datos nulos.

```
# Asignamos valores NA
titanicData$Cabin[sinCabina] <-NA

# Volvemos a comprobar si ahora disponemos de valores n
length(which(is.na(titanicData$Cabin)))
```

```
## [1] 40
```

Se puede observar que ahora sí diponemos de valores nulos.

Procedemos a comprobar con el resto de variables ya que podemos tener más valores NA ocultos.

```
# Comprobamos para las otras columnas
apply(X =titanicData[,c("Name","Sex","Ticket","Embarked")],
      MARGIN =2,
      FUN =function(x) length(which(x=="")))
```

```
##     Name      Sex   Ticket Embarked
##        0        0        0        2
```

Se puede observar que la variable embarked también dispone de 2 valores de strings vacíos.

Por lo que también convertiremos estos dos campos vacíos en valores NA.

```
# Ponemos valores NA para los datos vacíos
titanicData$Embarked[titanicData$Embarked==""] <-NA

# Volvemos a comprobar si ahora disponemos de valores na
length(which(is.na(titanicData$Embarked)))
```

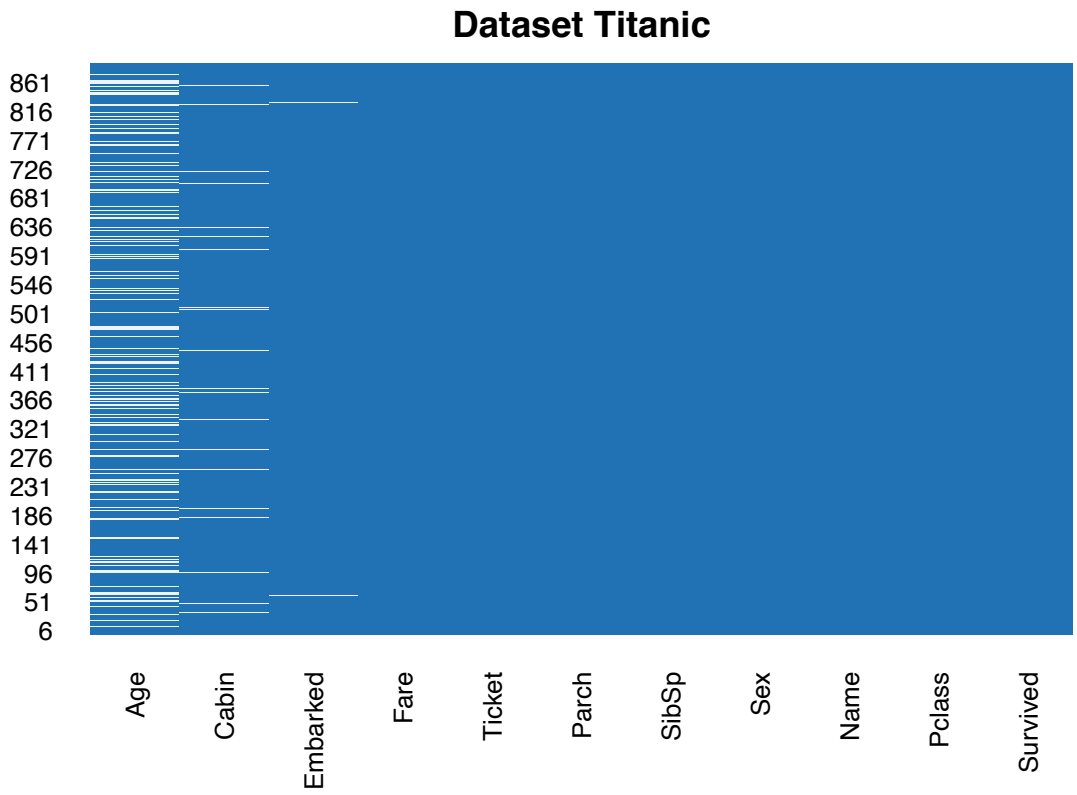
```
## [1] 2
```

Otro aspecto a tener en cuenta, de los 12 campos, el campo idPasajero no consideramos útil para el análisis en la siguiente sección.

```
# Borramos el campo PassengerId
titanicData$PassengerId <- NULL
```

Inicialmente solo teníamos valores NA en el campo edad.

```
missmap(obj =titanicData, main ="Dataset Titanic", legend =FALSE)
```



Ahora podemos observar en la gráfica que tenemos más valores NA's.

Se podrían omitir completamente los registros u observaciones que tengan valores vacíos o nulos usando la función na.omit(), pero gestionaremos estos valores de la siguiente manera:

Para gestionar los valores na del campo embarked comprobamos los posibles valores que tiene esta variable categorica.

```
# Comprobamos los valores unicos
unique(titanicData$Embarked)
```

```
## [1] S    C    Q    <NA>
## Levels:  C Q S
```

```
# Comprobamos la cantidad de pasajeros desde donde embarcaron
xtabs(~Embarked, data =titanicData)
```

```
## Embarked
##       C   Q   S
##   0 168  77 644
```

Tenemos 3 principales niveles desde donde se embarcó.

En esta situación podemos reemplazar los valores faltantes por la clase mayoritaria.

```
titanicData$Embarked[is.na(titanicData$Embarked)] <-'S'
```

Y para la situación de la variable edad se puede gestionar añadiendo la media de edad.

```
# Tomamos la media para valores na de la variable "Age"
titanicData$Age[is.na(titanicData$Age)] <- mean(titanicData$Age,na.rm=T)

# Volvemos a comprobar si ahora disponemos de valores na
length(which(is.na(titanicData$Age)))
```

```
## [1] 0
```

Con el objetivo de no perder información útil, se ha aplicado con cautela de tal manera de no introducir falsedad en los datos.

Para gestionar los valores nulos en la columna cabin, procedemos añadir la etiqueta/constante Desconocido, de esta manera no eliminaríamos la fila entera, y pondremos a las personas en una categoría de cabina Desconocida.

```
# Tomamos valor "Desconocido" para los valores vacíos de la variable "cabin"
titanicData$Cabin[is.na(titanicData$Cabin)] <- "Desconocido"
```

2.3.2 Valores extremos

En este punto vamos a identificar los valores extremos, si son lógicos o estamos ante posible errores que puedan dañar nuestros resultados finales.

En primer lugar, vamos a ver dónde conviene buscar estos valores de tipo numérico. Si recordamos las columnas del dataset:

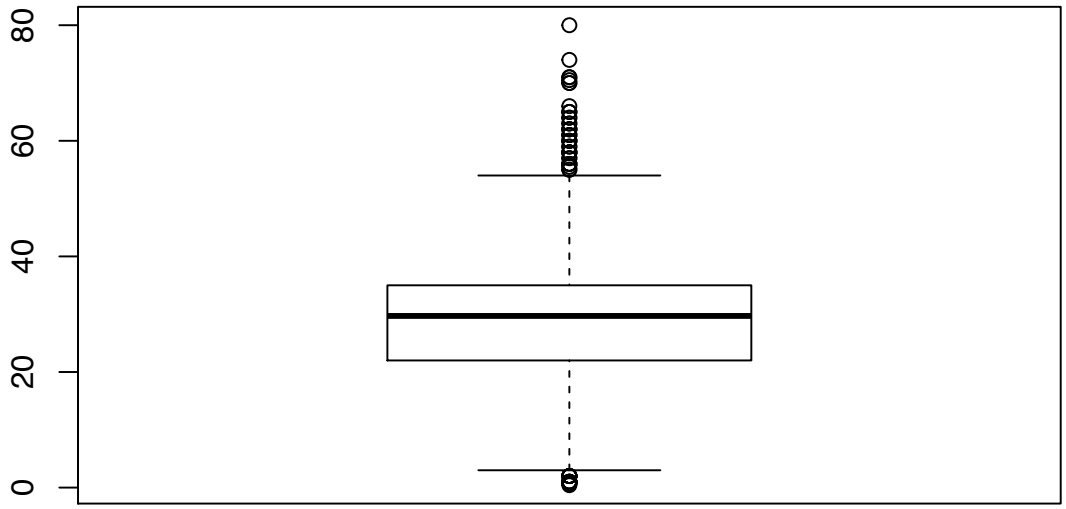
```
names(titanicData)
```

```
##  [1] "Survived" "Pclass"   "Name"     "Sex"      "Age"      "SibSp"
##  [7] "Parch"    "Ticket"   "Fare"     "Cabin"    "Embarked"
```

En estos casos, hemos ya visto en apartados anteriores cómo tratar los valores nulos o erróneos que saltarían a primera vista de verlos. Sin embargo, en el caso de los valores "Age" (edades) o "Fare" (tarifas), tenemos valores numéricos, que bien pueden ser no nulos o no erróneos, pueden tener valores absurdos que no sean correctos, o tan extremos que nos puedan hacer dudar de ser ciertos.

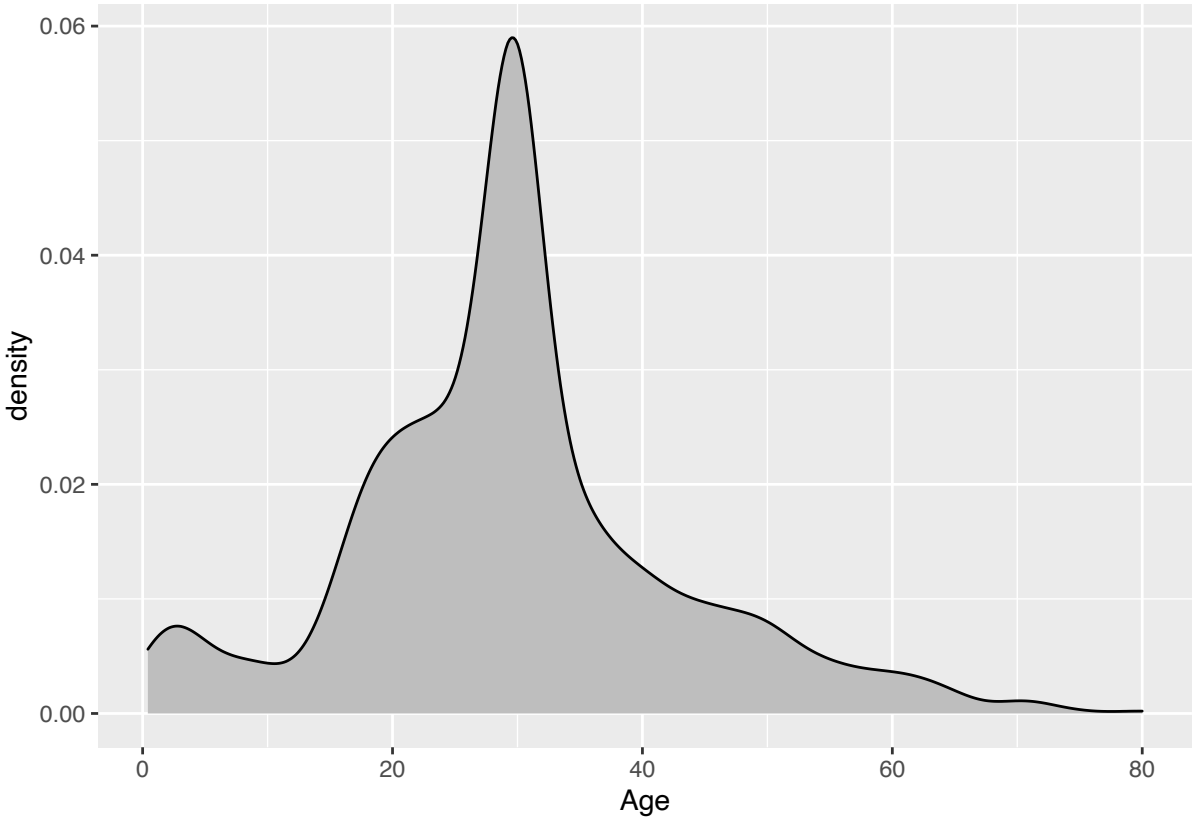
Así a priori, imaginemos que la edad de un pasajero que fuera de 500 años o bien una tarifa exageradamente alta. Así vamos a explorar estos datos a ver si encontramos algo que llame la atención:

```
# Pintamos el gráfico de outliers para el campo edad
boxplot(titanicData$Age)
```



Pintamos la gráfica de edades comprendidas.

```
# Cargamos las librerias necesarias
library(ggplot2)

# Pintamos el grafico de edades
ggplot(titanicData, aes(x = Age)) + geom_density(fill='gray')
```



Podemos ver que todos los valores están en rango que oscila entre 0 y 80 (tiene lógica tratándose de edades).

Por otro lado vamos a echar un vistazo a esos valores más altos en cuanto a edad.

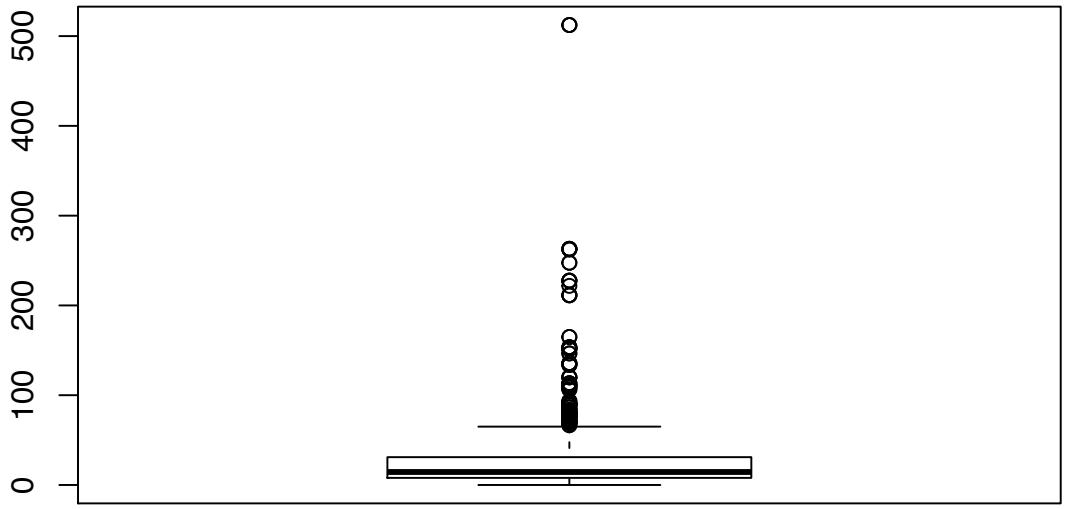
```
tail(sort(titanicData$Age),5)
```

```
## [1] 70.5 71.0 71.0 74.0 80.0
```

Como vemos, los valores más altos no son nada alarmante, aunque son edades desde luego altas para la época, los valores de personas por encima de los 70 como vemos son extremadamente bajos.

Por otro lado, vamos a las tarifas ("Fare"):

```
#Hacemos un boxplot de Fare
boxplot(titanicData$Fare)
```



Como vemos, en el caso de las tarifas, hay una concentración alta en la tarifas más bajas. Por otro lado, vemos unas cuantas tarifas más altas, estos valores llaman la atención, pero en concreto uno en el que puede apreciarse que se paga más de 500 de precio por el billete.

```
# Mostramos los 5 valores más elevados
tail(sort(titanicData$Fare),5)
```

```
## [1] 263.0000 263.0000 512.3292 512.3292 512.3292
```

Vemos que hay 4 personas en total que han pagado esa tarifa tan alta, además gracias a la exploración de los valores más altos, sabemos que han pagado exactamente el mismo precio 512.3292 (3 pasajeros), aunque el número es extraño a priori, esta claro que se trata que algunos pasajeros pagaron este valor elevado por el ticket.

2.4 Análisis de los datos

La selección de datos que se quiere comparar, son los datos de los pasajeros que sobrevivieron y los que murieron, además de las variables que tiene más influencia sobre estos dos aspectos, de esta manera podríamos obtener las conclusiones sobre qué grupo de personas estaba más expuesto a morir.

Primero de todo miraremos la cantidad de pasajeros que sobrevivieron:

```
  sum(titanicData$Survived==0)
```

```
## [1] 549
```

También miraremos los que murieron:

```
  sum(titanicData$Survived==1)
```

```
## [1] 342
```

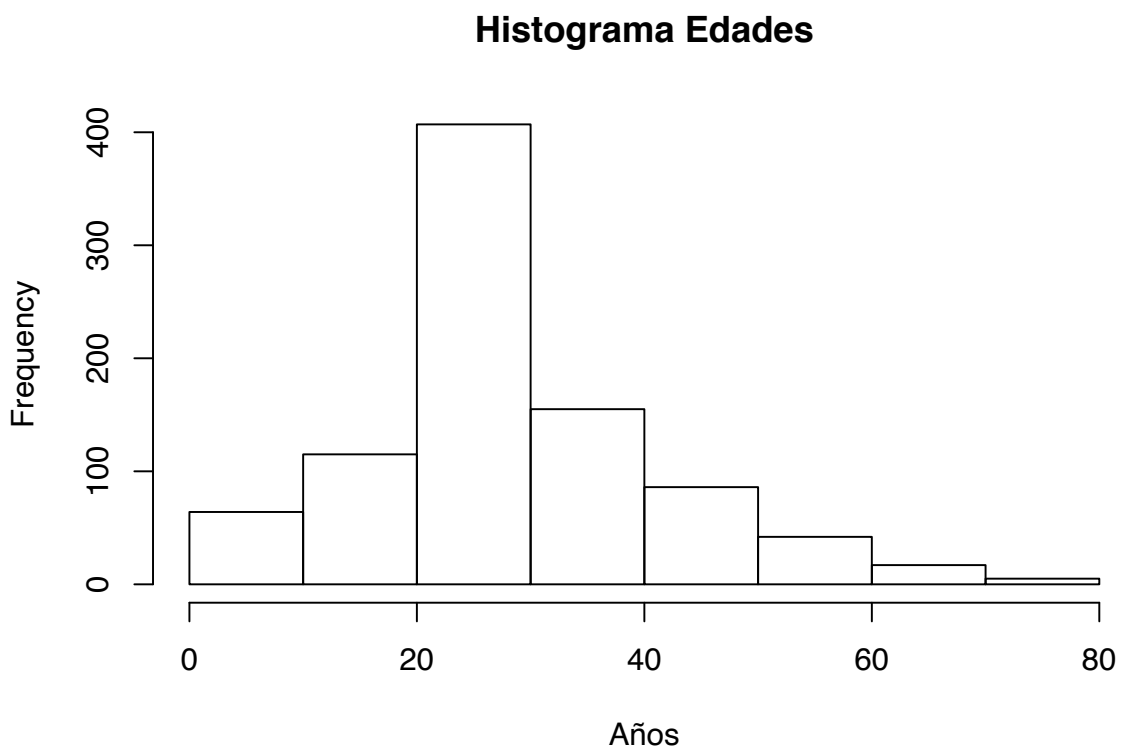
Miraremos la edad media de los pasajeros.

```
  mean(titanicData$Age, na.rm=TRUE)
```

```
## [1] 29.69912
```

Para el análisis consideramos que la variable edad puede tener importancia sobre los resultados, debido a que las personas mayores y niños son un grupo vulnerable, para el cual sería más dificíl una evacuación.

```
# Pintamos el histograma de frecuencia de edades
hist(titanicData$Age, main = "Histograma Edades", xlab = "Años")
```



Podemos ver el gráfico de frecuencia, la media de edades rondaba sobre los 20-30 años de edad.

Pero continuando con el análisis nos interesa saber que rango de edad sobrevivieron/murieron más personas.

```
# Guardamos en las dos variables los datos correspondientes
sobrevivientes_edad <- titanicData[which(titanicData$Survived==1),"Age"]
muertes_edad <- titanicData[which(titanicData$Survived==0),"Age"]

# Imprimimos por pantalla la variables anteriores
print (mean(sobrevivientes_edad))
```

```
## [1] 28.54978
```

```
print (mean(muertes_edad))
```

```
## [1] 30.4151
```

Vemos la media del grupo de sobrevivientes es de 28.5 años mientras que la media del grupo de fallecidos es de 30.4 años. En el siguiente a partado podremos ver estos datos visuales de las personas que sobrevivieron y los que NO en un gráfico boxplot.

La homogeneidad va referida a si son parecidas las varianzas entre muestras. Para este caso, vamos brevemente a aplicar la prueba de Bartlett's, que, aunque funciona mejor con distribuciones normales de probabilidad, puede ser útil en este caso: Se seleccionarán la edad (Age) y el embarque (Embarked).

```
bartlett.test(titanicData$Age, titanicData $Embarked)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  titanicData$Age and titanicData$Embarked
## Bartlett's K-squared = 8.9798, df = 2, p-value = 0.01122
```

El método prueba la hipótesis nula de que la varianza entre muestras es igual. Aquí por ejemplo al ser el valor p=0.01122 inferior al valor crítico 0,05 que se suele considerar de nivel de significancia (es decir en un rango del 95%), se puede concluir que la varianza entre muestras es parecida en comparación a las demás muestras presentes.

A continuación miraremos con que variables tiene sentido realizar un proceso de discretización.

```
# ¿Con qué variables tendría sentido un proceso de discretización?
apply(titanicData,2, function(x) length(unique(x)))
```

```
## Survived   Pclass     Name      Sex      Age    SibSp    Parch   Ticket
##        2        3      891        2       89        7        7      681
##     Fare    Cabin Embarked
##      248      149        3
```

Discretizamos las variables con pocas clases.

```
# Procedemos a seleccionar las columnas con pocas clases
cols<-c("Survived","Pclass","Sex","Embarked")

# Las pasamos a factor
for (i in cols){
  titanicData[,i] <- as.factor(titanicData[,i])
}
```

Después de los cambios, analizamos la nueva estructura del conjunto de datos

```
# Se Muestra estructura del conjunto de datos
str(titanicData)
```

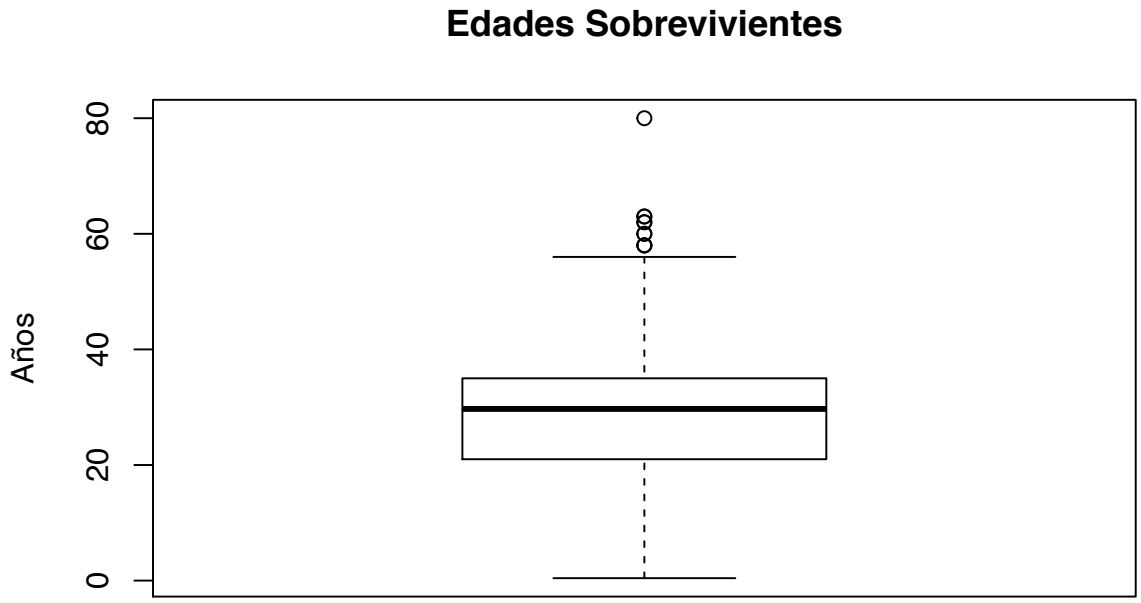
```
## 'data.frame':    891 obs. of  11 variables:
##  $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name    : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 .
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 ...
##  $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch   : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket  : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin   : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

A continuación se procede a la representación las relaciones entre las diferentes variables para obtener resultados visuales.

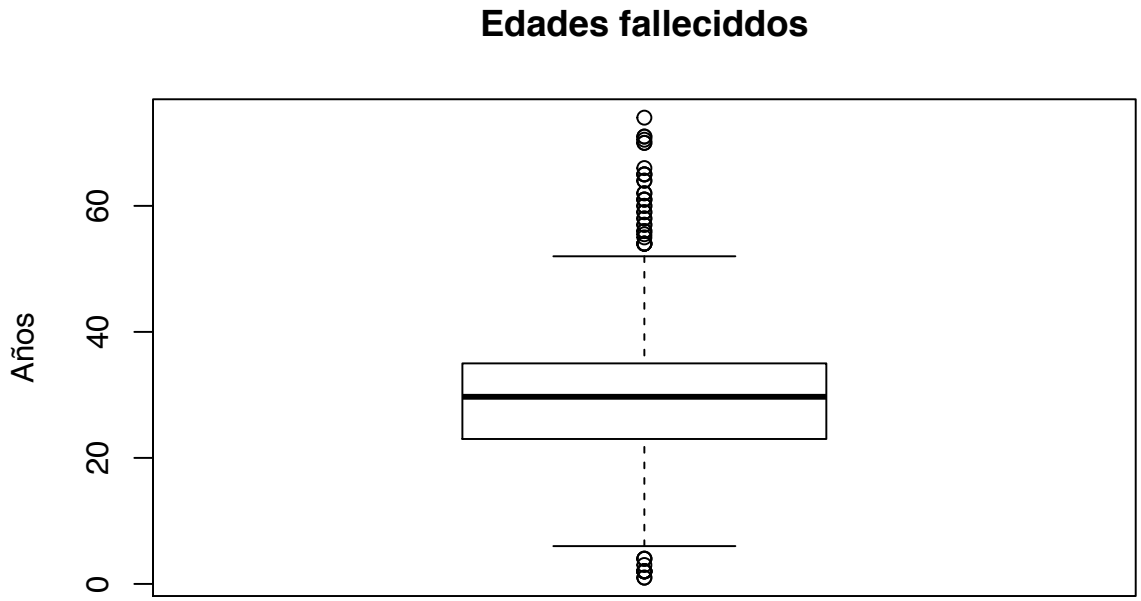
2.5 Representación de los resultados

Primero representaremos los sobrevivientes y los fallecidos en gráficos boxplot.

```
# Pintamos el grafico de edades sobrevivientes
boxplot(sobrevivientes_edad, main ="Edades Sobrevivientes", ylab ="Años")
```




```
# Pintamos el grafico de edades que fallecieron
boxplot(muertes_edad,main ="Edades falleciddos", ylab ="Años")
```

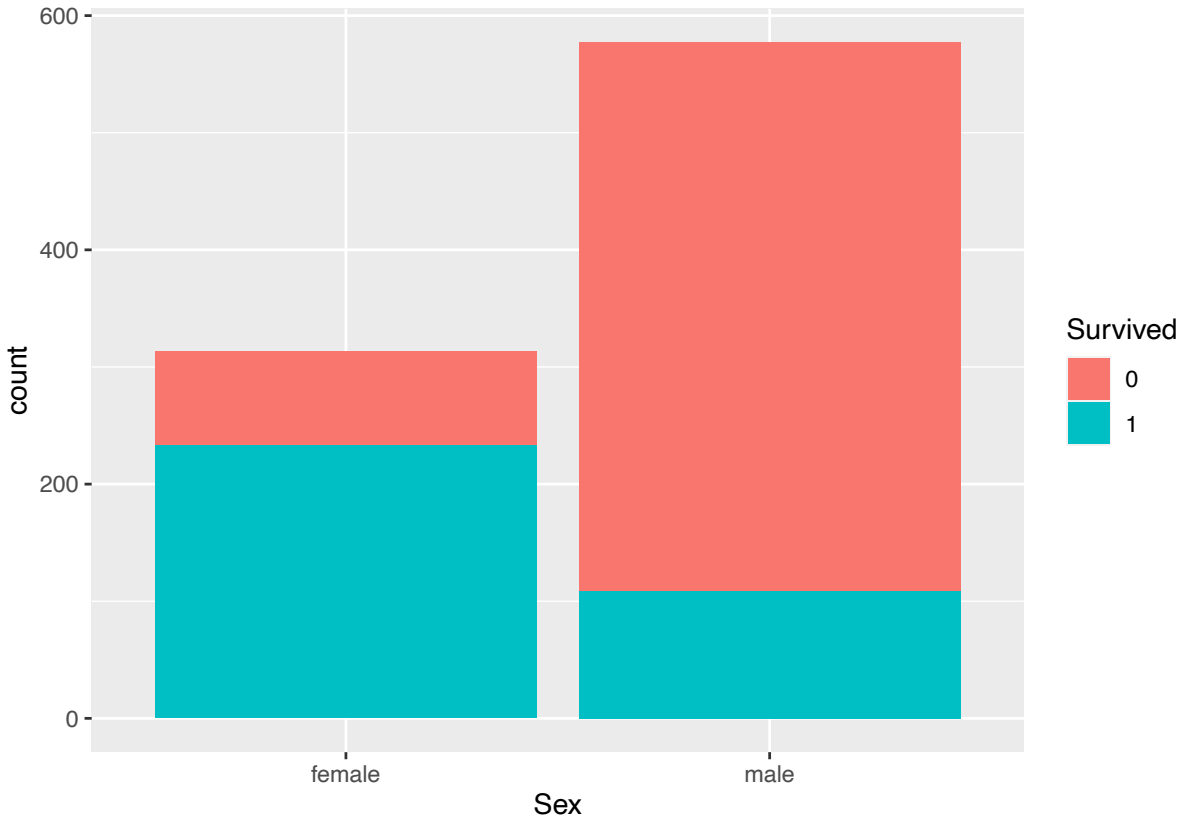


Una de las observaciones más importantes de estos gráficos es que mientras más joven era la persona más probabilidad tenía de sobrevivir del hundimiento, su media calculada en el apartado anterior se encontraba sobre los 28 años mientras que para las personas que murieron se encontraba sobre los 30 años.

En el segundo gráfico se puede ver claramente los valores extremos son de personas mayores.

Se procede a comprobar la relación entre las variable sexo con la de supervivencia.

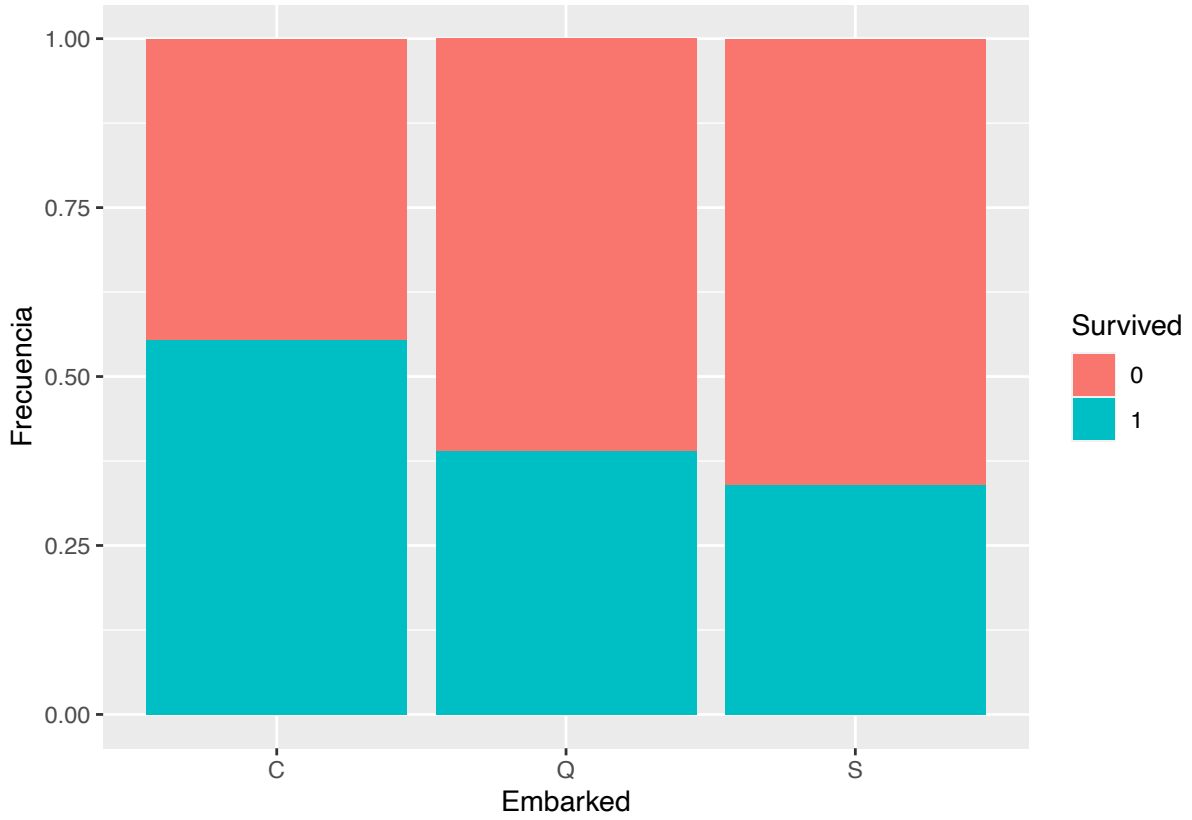
```
# Importamos las librerías necesarias par alos gráficos
library(ggplot2)
library(dplyr)

# Guardamos el número de filas en una variable
filas=dim(titanicData)[1]

# Visualizamos la relación entre las variables "Sex" y "Survived":
ggplot(data=titanicData[1:filas,],aes(x=Sex,fill=Survived))+geom_bar()
```



Los datos con respecto al género no estan balanceados, pero si observamos el gráfico podemos obtener que los hombres podrían llegar a sobrevivir con mayor probabilidad que las mujeres.

Se comprueba la relación entre lugar de embarque con la supervivencia.

```
# Se pinta el gráfico
ggplot(data = titanicData[1:filas,],aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+ylab("Frecu
```



En esta gráfica podemos ver los puertos de embarque y los porcentajes de supervivencia en funcion del puerto.

Pintaremos una matriz con los porcentajes de frecuencia del gráfico anterior para poder interpretar mejor los resultados.

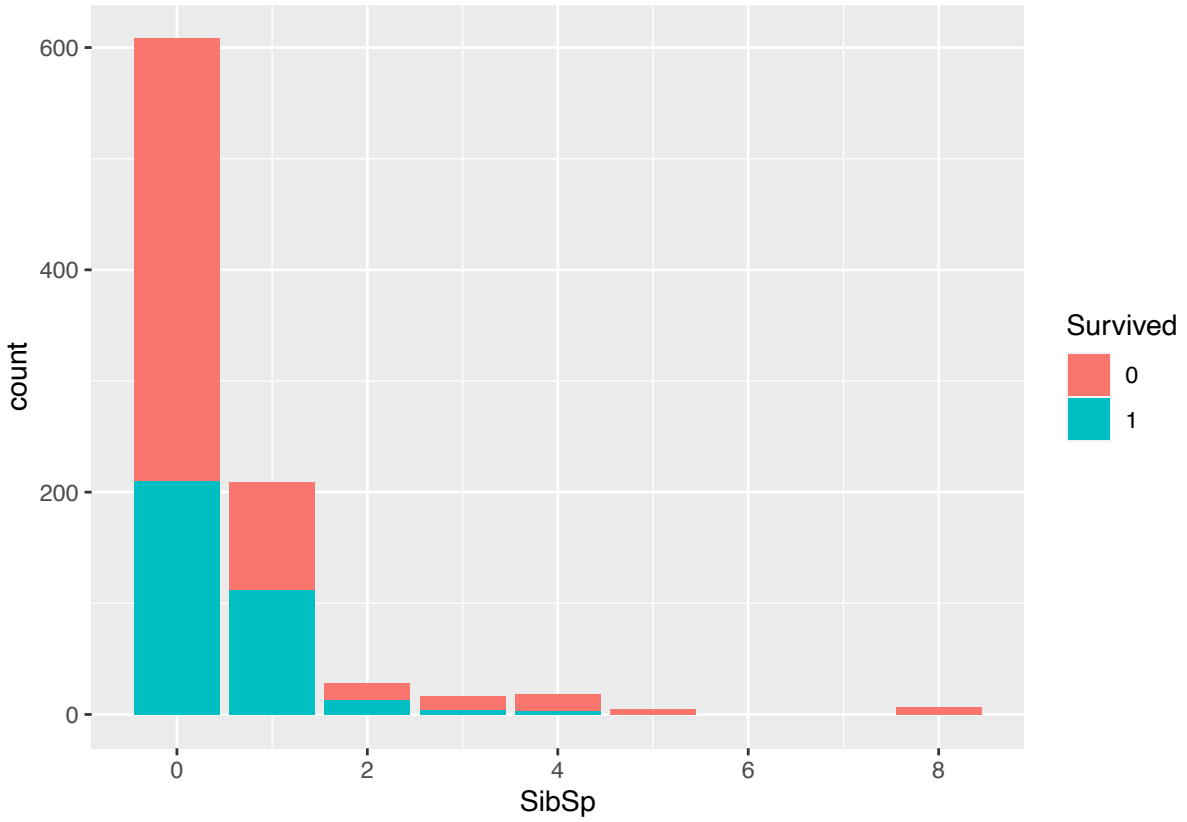
```
t<-table(titanicData[1:filas,]$Embarked,titanicData[1:filas,]$Survived)
for (i in 1:dim(t)[1]){
    t[i,]<-t[i,]/sum(t[i,])*100
}
t
```

```
##
##            0        1
##
##   C 44.64286 55.35714
##   Q 61.03896 38.96104
##   S 66.09907 33.90093
```

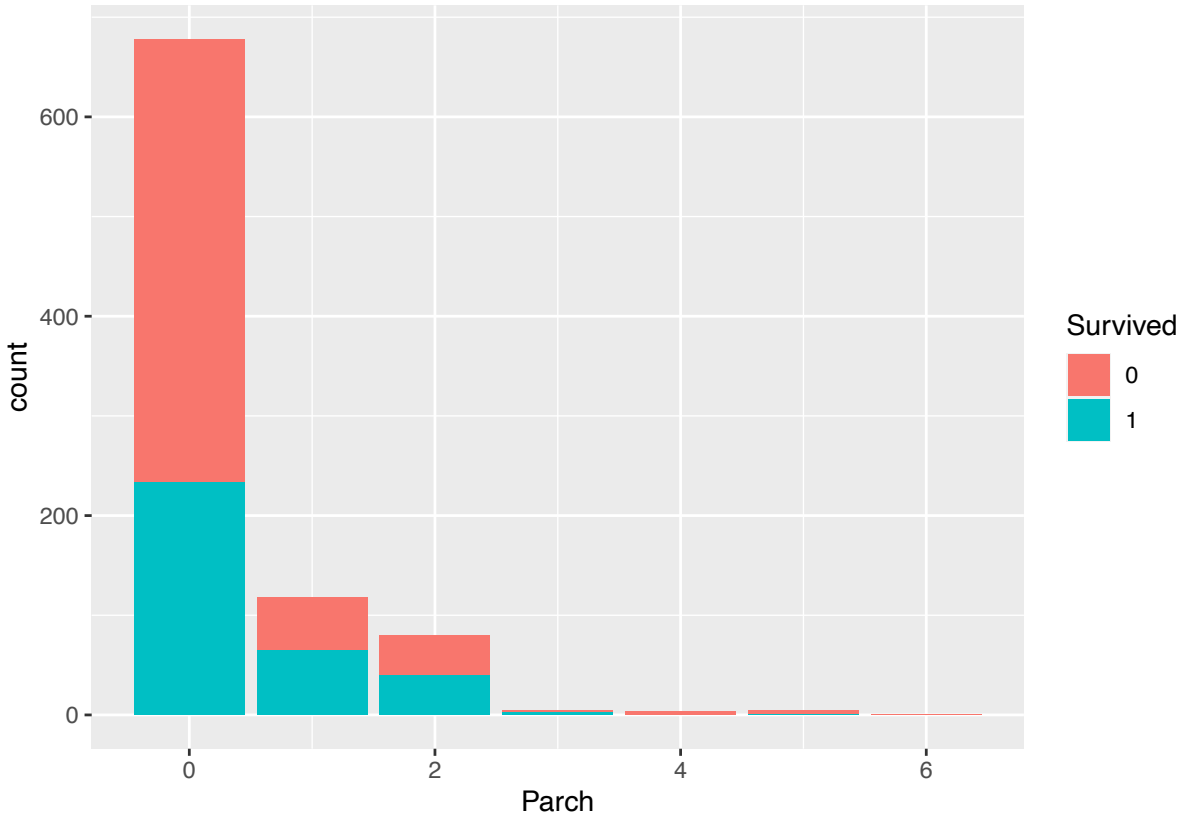
De esta matriz de porcentaje de frecuencia se extrae que en el puerto Q la probabilidad de sobrevivir es de un 61,03%.

Compararemos ahora dos gráficos de frencuencias: Survived-SibSp y Survived-Parch

```
# Survivial como función de SibSp y Parch
ggplot(data = titanicData[1:filas,],aes(x=SibSp,fill=Survived))+geom_bar()
```




```
ggplot(data = titanicData[1:filas,],aes(x=Parch,fill=Survived))+geom_bar()
```



Vemos como las forma de estos dos gráficos es similar. Este hecho nos puede indicar presencia de correlaciones altas, esto nos quiere decir que podrían a mayor numéro de hermanos o familiarias que viajaban menor era la probabilidad de sobrevivir.

2.6 Resolución - Conclusión

- Uno de los datos más importantes obtenidos en los resultados de esta practica, es que mientras más joven era la personas mayor era la probabilidad de sobrevivir del hundimiento mientras que las personas mayores tenían menos probabilidad de sobrevivir.
- Los pasajeros que embarcaron en Q tenían más probabilidad de sobrevivir, puede que se deba a la cercanía de los salvavidas.
- A pesar de que los datos no estan complementamente balanceados en cuanto al género, se puede decir que los hombres tenían mayor probabilidad de sobrevivir al hundimiento.
- Mientras mayor era el número de familiares durante el viaje menor era la probabilidad de sobrevivir.

3 Contribuciones

Contribución	Firma
Investigación previa	AGG, ETP
Redacción de las respuestas	AGG, ETP
Desarrollo código	AGG, ETP