

Tipología y Ciclo de Vida de los Datos: Práctica 2

Autor: Francisco José Núñez Sánchez-Agustino

Junio 2021

- Descripción del conjunto de datos
- Importación de las librerías necesarias
- Integración y selección de los datos de interés
- Limpieza de los datos
- Análisis de los datos
- Conclusiones

Ruta al repositorio de la práctica

<https://github.com/fnunezs/data-analysis> (<https://github.com/fnunezs/data-analysis>)

Descripción del conjunto de datos

El conjunto de datos seleccionado para esta práctica contiene parte de la lista de pasajeros que viajaban a bordo del buque *RMS Titanic* en su viaje inaugural entre Southampton y Nueva York, que en la noche del 14 al 15 de abril de 1912 se hundió frente a las costas de Terranova después de colisionar con un iceberg. El impacto con la masa de hielo ocasionó una vía de agua por debajo de la línea de flotación del casco que provocó el hundimiento del barco en menos de tres horas.

Este naufragio es una de las mayores catástrofes marítimas de la historia en tiempos de paz, en la que perecieron un total de 1.496 personas por golpes, caídas, ahogamiento e hipotermia. Esta tragedia supuso una gran commoción a nivel mundial en su época, ya que se trataba del buque de pasajeros más grande y lujoso construido hasta la fecha. Las conclusiones de las investigaciones llevadas a cabo tras su hundimiento sirvieron para desarrollar nuevos reglamentos que mejoraron la seguridad marítima en los años posteriores.

Los restos del *Titanic* fueron descubiertos en 1985 por el oceanógrafo Robert Ballard en el fondo del Atlántico Norte a casi 4.000 metros de profundidad. Desde el descubrimiento del pecio han sido recuperados numerosos objetos del buque que están expuestos en diferentes museos.

En lo que respecta al origen del *dataset*, este contiene parte de la lista de pasajeros del *Titanic* recopilada por Michael A. Findlay y publicada por la organización sin ánimo de lucro *Encyclopedia Titanica* (<https://www.encyclopedia-titanica.org/>), fundada en 1996 por Philip Hind para investigar todo lo referente al *Titanic* y a su hundimiento. Su sitio web ofrece, entre otros datos, biografías de la tripulación y sus pasajeros, planos del barco, así como artículos de historiadores e investigadores de este naufragio.

Este es uno de los *dataset* más populares para iniciarse en técnicas de *Machine Learning* y está especialmente indicado para analizar algoritmos de clasificación, ya que para cada pasajero indica si este sobrevivió o no al naufragio, permitiendo analizar la relación entre este hecho y características como el sexo, edad o la clase en la que viajaba.

Importación de las librerías necesarias

Una vez descrito el conjunto de datos vamos a importar las librerías del lenguaje R que utilizaremos para analizarlo:

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('ggcorrplot')) install.packages('ggcorrplot'); library('ggcorrplot')
if (!require(ggpubr)) install.packages('ggpubr'); library(ggpubr)
if (!require(grid)) install.packages('grid'); library(grid)
if (!require(gridExtra)) install.packages('gridExtra'); library(gridExtra)
if (!require('tidyverse')) install.packages('tidyverse'); library('tidyverse')
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('purrr')) install.packages('purrr'); library('purrr')
if (!require('nortest')) install.packages('nortest'); library('nortest')
```

Integración y selección de los datos de interés

En primer lugar vamos a cargar los ficheros que componen este conjunto de datos. Estos han sido descargados de una de las competiciones más populares del sitio web Kaggle (<https://www.kaggle.com/c/titanic/data>), que consiste en crear un modelo de aprendizaje automático para predecir qué pasajeros sobrevivieron a este naufragio, siendo este el motivo por el que el *dataset* se presenta como dos ficheros independientes: uno para “entrenar” el algoritmo clasificador (*train.csv*) y otro para comprobar su capacidad predictiva (*test.csv*).

```
# Carga de Los ficheros del conjunto de datos
titanic_train <- read.csv('train.csv', stringsAsFactors = FALSE)
titanic_test <- read.csv('test.csv', stringsAsFactors = FALSE)
```

A continuación vamos a comprobar si los datos cargados de estos ficheros tienen la misma estructura, ya que la intención es combinarlos para generar un único conjunto de datos:

```
# Estructura del conjunto de datos "train"
str(titanic_train)
```

```

## '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)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
## $ 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" ...

```

```

# Estructura del conjunto de datos "test"
str(titanic_test)

```

```

## 'data.frame': 418 obs. of 11 variables:
## $ PassengerId: int 892 893 894 895 896 897 898 899 900 901 ...
## $ Pclass     : int 3 3 2 3 3 3 3 2 3 3 ...
## $ Name       : chr "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Thomas Francis" "Wirz, Mr. Albert" ...
## $ Sex        : chr "male" "female" "male" "male" ...
## $ Age        : num 34.5 47 62 27 22 14 30 26 18 21 ...
## $ SibSp      : int 0 1 0 0 1 0 0 1 0 2 ...
## $ Parch      : int 0 0 0 0 1 0 0 1 0 0 ...
## $ Ticket     : chr "330911" "363272" "240276" "315154" ...
## $ Fare        : num 7.83 7 9.69 8.66 12.29 ...
## $ Cabin      : chr "" "" "" ...
## $ Embarked   : chr "Q" "S" "Q" "S" ...

```

La ejecución del código anterior muestra que el conjunto de entrenamiento (*titanic_train*) tiene un total de 891 observaciones con 12 variables cada una, mientras que el de prueba (*titanic_test*) contiene 418 con 11 variables cada una. Analizando la lista de variables obtenida podemos observar que la diferencia entre ambos se debe a que en el de prueba se ha excluido el atributo que indica si el pasajero sobrevivió al naufragio (*Survived*).

Al tratarse de la variable que da sentido a este conjunto de datos y sobre la que basaremos el análisis, vamos a intentar obtenerla a partir de otra copia de este conjunto de datos que contenga esta información para todos los pasajeros.

En concreto, vamos a utilizar otra versión (https://www.kaggle.com/heptapod/titanic?select=train_and_test2.csv) del mismo *dataset* que también está disponible en la web de Kaggle. De este conjunto de datos vamos a extraer el valor de la variable *Survived* de los pasajeros cuyo identificador numérico *PassengerId* coincide con el que tiene en el *dataset* de prueba (*titanic_test*), añadiendo la información extraída a cada pasajero de este conjunto de datos.

```

# Cargamos el conjunto de datos que contiene la variable "Survived" para todos los pasajeros
titanic_all <- read.csv('train_and_test2.csv', stringsAsFactors = FALSE)

# Cambiamos el nombre de la columna "2urvived" por "Survived" en el conjunto que acabamos de cargar
colnames(titanic_all)[28] <- "Survived"

# Añadimos la variable "Survived" a cada pasajero del conjunto de prueba según su identificador.
titanic_test$Survived <- titanic_all$Survived[titanic_all$PassengerId %in% titanic_test$PassengerId]

# Colocamos la columna en la misma posición que se encuentra en el conjunto de datos de entrenamiento
titanic_test <- titanic_test %>% relocate(Survived, .before = Pclass)

```

Ahora que tenemos los conjuntos de entrenamiento y prueba con la misma estructura vamos a combinar sus datos para generar un único conjunto de datos:

```
titanic_data <- rbind(titanic_train, titanic_test)
```

A continuación vamos a comprobar que el nuevo conjunto de datos generado no contiene registros duplicados:

```
anyDuplicated(titanic_data)
```

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

El resultado anterior muestra que no existe ninguna observación duplicada en el *dataset* que hemos generado. El siguiente código muestra cual es su estructura:

```
str(titanic_data)
```

```

## 'data.frame': 1309 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 (Florenc e Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Pee l)" ...
## $ 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" ...

```

Como era de esperar, tenemos un total de **1.309 observaciones** con **12 variables** que se corresponde con la suma de las del conjunto de entrenamiento (891) y las de prueba (418).

La descripción de cada una de estas variables es la siguiente:

- *PassengerId*: identificador numérico del pasajero.
- *Survived*: variable categórica que identifica si el pasajero sobrevivió (valor “1”) o no (valor “0”). Se trata de la variable que clasifica las observaciones de este *dataset*.
- *Pclass*: variable categórica que indica la clase en la que viajaba el pasajero. Los valores son “1” para la primera y más lujosa, “2” para la segunda y “3” para la tercera clase que era la más económica.
- *Name*: variable de texto que contiene el nombre del pasajero, incluyendo también la fórmula de tratamiento (Mr, Miss, etc.)
- *Sex*: variable categórica que indica el sexo del pasajero. Sus valores son “*male*” para los hombres y “*female*” para las mujeres.
- *Age*: valor numérico que representa la edad del pasajero en años. Contiene decimales cuando se trata de menores de un año o es un valor estimado.
- *SibSp*: variable numérica que contiene el número de familiares adultos del pasajero que se encontraban a bordo (hermanos, cónyuge, etc.).
- *Parch*: variable numérica que, en el caso de los niños contiene el número de progenitores que se encontraban a bordo, y en el caso de los adultos el número de hijos que viajaban con él.
- *Ticket*: variable de texto que contiene el identificador del billete del pasajero.
- *Fare*: valor numérico decimal que representa el precio del billete pagado por el pasajero.
- *Cabin*: número de camarote del pasajero.
- *Embarked*: puerto en el que embarcó el pasajero. Sus valores son “C” para la ciudad francesa de Cherbourg, “Q” para Queenstown en Irlanda (actualmente Cobh) y “S” para Southampton en Inglaterra.

Una vez conocido el significado y características de estas variables vamos a crear un par de **atributos derivados** de los anteriores que serán de utilidad en el análisis que se realizará posteriormente:

- El primero contendrá el **tamaño de la familia** embarcada del pasajero. Este dato lo obtendremos a partir de las variables *SibSp* y *Parch*, tal y como muestra el siguiente código:

```
# Calculamos el tamaño de la familia embarcada del pasajero
titanic_data$FamilySize <- titanic_data$Parch + titanic_data$SibSp + 1
```

- El segundo atributo que generaremos será el tipo de **tratamiento del pasajero**. Este dato será útil para determinar su clase social e incluso su edad, ya que en la época en la que se produjo el hundimiento los niños recibían el tratamiento formal de “*Master*” ([https://en.wikipedia.org/wiki/Master_\(form_of_address\)](https://en.wikipedia.org/wiki/Master_(form_of_address))). El siguiente código generará esta variable a partir del nombre del pasajero (*Name*):

```

# Extraemos el tratamiento del nombre del pasajero (Mr, Mrs, Miss, etc.)
titanic_data$title <- sapply(titanic_data$name, FUN=function(x) {strsplit(x, split='[,.]')[[1]][2]})

# Eliminamos el espacio en blanco al inicio del título
titanic_data$title <- sub(' ', '', titanic_data$title)

# Agrupamos militares, médicos y clérigos como "Officer"
titanic_data$title[titanic_data$title %in% c('Capt', 'Col', 'Major', 'Dr', 'Rev')] <- 'Officer'

# Agrupamos pasajeros con título nobiliario como "Nobility"
titanic_data$title[titanic_data$title %in% c('Jonkheer', 'Sir', 'the Countess', 'Lady')] <- 'Nobility'

# Agrupamos a Los hombres que no pertenecen a Los anteriores como 'Mr'
titanic_data$title[titanic_data$title %in% c('Mr', 'Don')] <- 'Mr'

# Agrupamos a Las mujeres casadas que no pertenecen a Los anteriores como 'Mrs'
titanic_data$title[titanic_data$title %in% c('Mrs', 'Mme', 'Ms', 'Dona')] <- 'Mrs'

# Agrupamos a Las mujeres solteras que no pertenecen a Los anteriores como 'Miss'
titanic_data$title[titanic_data$title %in% c('Miss', 'Mlle')] <- 'Miss'

# Mostramos Los distintos grupos de tratamiento que hemos creado
unique(titanic_data$title)

```

```
## [1] "Mr"      "Mrs"     "Miss"    "Master"   "Officer" "Nobility"
```

Limpieza de los datos

La mayoría de las variables incluidas en el *dataset* parecen útiles para intentar comprender los factores que influyeron en la supervivencia de los pasajeros del Titanic.

Las únicas variables que no parece que vayan a ser de utilidad para este objetivo son el identificador del pasajero *PassengerId* - que resultó fundamental para llevar a cabo la integración de datos - y el identificador o número de billete del pasajero (*Ticket*).

El siguiente código elimina estas dos variables del conjunto de datos:

```

# Eliminamos del dataset la variable PassengerId
titanic_data <- subset(titanic_data, select = -PassengerId)

# Eliminamos del dataset la variable Ticket
titanic_data <- subset(titanic_data, select = -Ticket)

```

Por otro lado, la creación de la variable *FamilySize*, que combina los valores de *SibSp* y *Parch*, *a priori* permite prescindir de estas variables. El código que se muestra a continuación las elimina también de este *dataset*:

```
# Eliminamos del dataset la variable SibSp
titanic_data <- subset(titanic_data, select = -SibSp)

# Eliminamos del dataset la variable Parch
titanic_data <- subset(titanic_data, select = -Parch)
```

Para facilitar el análisis que llevaremos a cabo, vamos a sustituir los valores de las variables categóricas *Survived* y *Pclass* por cadenas de texto. Para la primera reemplazaremos los valores “1” y “0” por “Yes” y “No”, en el caso de la segunda convertiremos sus valores numéricos a texto. El siguiente código realiza estas operaciones:

```
# Sustituimos los valores "1" y "0" de la variable Survived por "Yes" y "No"
titanic_data$Survived[titanic_data$Survived == "1"] <- "Yes"
titanic_data$Survived[titanic_data$Survived == "0"] <- "No"

# Convertimos los valores de la variable Pclass a texto
titanic_data$Pclass <- as.character(titanic_data$Pclass)
```

Tratamiento de los valores nulos y desconocidos

A continuación, vamos a comprobar si existen valores nulos o perdidos en alguna de las variables. Para ello vamos a obtener las estadísticas básicas del *dataset*:

```
summary(titanic_data)
```

```

##   Survived          Pclass          Name          Sex
## Length:1309      Length:1309      Length:1309      Length:1309
## Class :character Class :character Class :character Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
## 
## 
## 
##   Age            Fare          Cabin          Embarked
## Min.   : 0.17    Min.   : 0.000  Length:1309      Length:1309
## 1st Qu.:21.00   1st Qu.: 7.896  Class :character  Class :character
## Median :28.00   Median : 14.454  Mode   :character  Mode   :character
## Mean   :29.88   Mean   : 33.295
## 3rd Qu.:39.00   3rd Qu.: 31.275
## Max.   :80.00   Max.   :512.329
## NA's    :263    NA's    :1
##   FamilySize      Title
## Min.   : 1.000  Length:1309
## 1st Qu.: 1.000  Class :character
## Median : 1.000  Mode  :character
## Mean   : 1.884
## 3rd Qu.: 2.000
## Max.   :11.000
##

```

Analizando estos resultados no parece que existan valores nulos en ninguna de las **variables categóricas**. Para asegurar que tampoco contienen valores desconocidos vamos a ejecutar el siguiente código, que muestra el total de cadenas de texto vacías presentes entre sus valores:

```

# Localizamos las variables categóricas
variables_categoricas <- unlist(lapply(titanic_data,is.character))

# Las seleccionamos en el conjunto de datos
variables_seleccionadas <- titanic_data[ , variables_categoricas]

# Contamos el total de cadenas de texto vacías presentes en cada variable
variables_con_valores_desconocidos <- sapply(variables_seleccionadas, function(x) sum(x == ''))

# Mostramos los resultados
variables_con_valores_desconocidos

```

Survived	Pclass	Name	Sex	Cabin	Embarked	Title
0	0	0	0	1014	2	0

El resultado anterior muestra que existen valores desconocidos para las variables *Cabin* y *Embarked*.

En el caso de la primera, estos valores están presentes en el 77,46 % de los registros, motivo por el que vamos a prescindir de esta variable, ya que el escaso volumen de información que aporta no parece significativo para determinar la supervivencia del pasajero.

Respecto al puerto de embarque, sustituiremos los dos valores con cadenas de texto vacías por el carácter “U” para indicar que se trata de un valor desconocido (“Unknown”).

El siguiente código realiza estas operaciones:

```
# Eliminamos la variable que contiene el número de camarote (Cabin)
titanic_data <- subset(titanic_data, select = -Cabin)

# Reemplazamos las cadenas vacías de la variable "Embarked" por el texto "U"
titanic_data$Embarked[titanic_data$Embarked == '')] <- 'U'
```

En lo que respecta a las **variables numéricas**, el siguiente código pone de manifiesto que existen valores nulos en las variables *Age* y *Fare*:

```
# Comprobamos si existen valores nulos en el dataset
colSums(is.na(titanic_data))
```

```
##   Survived      Pclass       Name     Sex     Age     Fare Embarked
##       0          0          0        0     263       1         0
## FamilySize      Title
##       0          0
```

En el resultado anterior se observa que existen 263 pasajeros para los que se desconoce su edad, es decir, un 20% del total de observaciones.

Para evitar perder información útil no eliminaremos estos registros del *dataset*. En su lugar vamos a reemplazar los valores nulos por la media de edad del grupo al que pertenezca el pasajero, considerando estos tres grupos: hombres, mujeres y niños.

El criterio que seguiremos será el siguiente:

- Si es un hombre y su tratamiento (*Title*) es distinto de “Master” remplazaremos el valor nulo por la media de edad de los hombres embarcados.
- Si es una mujer y su tratamiento es distinto de “Master” remplazaremos el valor nulo por la media de edad de las mujeres embarcadas.
- Si tiene tratamiento de “Master” asumiremos que es un niño y sustituiremos el valor nulo por la media de edad de los niños embarcados.

El siguiente código realiza las operaciones que acabamos de describir:

```
# Seleccionamos los pasajeros hombres y calculamos su media de edad
hombres_titanic <- titanic_data %>% filter(Age > 17, Sex == 'male')
media_edad_hombres <- mean(hombres_titanic$Age, na.rm = TRUE)
media_edad_hombres <- round(media_edad_hombres, digits = 0)
media_edad_hombres
```

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

```
# Seleccionamos los pasajeros mujeres y calculamos su media de edad
mujeres_titanic <- titanic_data %>% filter(Age > 17, Sex == 'female')
media_edad_mujeres <- mean(mujeres_titanic$Age, na.rm = TRUE)
media_edad_mujeres <- round(media_edad_mujeres, digits = 0)
media_edad_mujeres
```

```
## [1] 33
```

```
# Seleccionamos los pasajeros niños y calculamos su media de edad
ninos_titanic <- titanic_data %>% filter(Age < 17)
media_edad_ninos <- mean(ninos_titanic$Age, na.rm = TRUE)
media_edad_ninos <- round(media_edad_ninos, digits = 0)
media_edad_ninos
```

```
## [1] 8
```

```
# Reemplazamos los valores nulos
titanic_data <- titanic_data %>% mutate(Age = case_when(
  is.na(Age) & Sex == 'male' & Title != 'Master' ~ media_edad_hombres,
  is.na(Age) & Sex == 'female' & Title != 'Master' ~ media_edad_mujeres,
  is.na(Age) & Title == 'Master' ~ media_edad_ninos,
  TRUE ~ Age
))
```

En el resultado de la ejecución del código que mostraba los valores nulos también se observaba la presencia de uno de estos valores para la variable que guarda el precio del billete (*Fare*). En este caso la estrategia que seguiremos será sustituirlo por la mediana.

El motivo para elegir este estimador en lugar de la media es la gran diferencia que se observa entre el valor máximo y el tercer cuartil de esta variable en las estadísticas generadas anteriormente, lo que podría deberse a valores “extremos”. Así pues, escogeremos la mediana al ser un estimador que no se verá influenciado por estos valores, al contrario que le ocurre a la media.

El siguiente código realiza esta operación:

```
# Calculamos la mediana de la variable Fare
mediana_Fare <- median(titanic_data$Fare, na.rm = TRUE)

# Reemplazamos los valores nulos de la variable Fare por la mediana
titanic_data$Fare[is.na(titanic_data$Fare)] <- mediana_Fare
```

Para garantizar que se han reemplazado todos los valores nulos que hemos detectado ejecutaremos el siguiente código:

```
# Comprobamos que ya no existen valores nulos
colSums(is.na(titanic_data))
```

```

##   Survived      Pclass       Name       Sex       Age      Fare     Embarked
##       0          0          0          0          0          0          0          0
## FamilySize      Title
##       0          0

```

El resultado anterior confirma que ya no existen valores nulos.

Tratamiento de los valores extremos

A continuación vamos a generar los gráficos de cajas (*boxplots*) para detectar si existen valores atípicos u *outliers* en las variables numéricas de este conjunto de datos:

```

# Generamos Las gráficas boxPlot de cada variable numérica

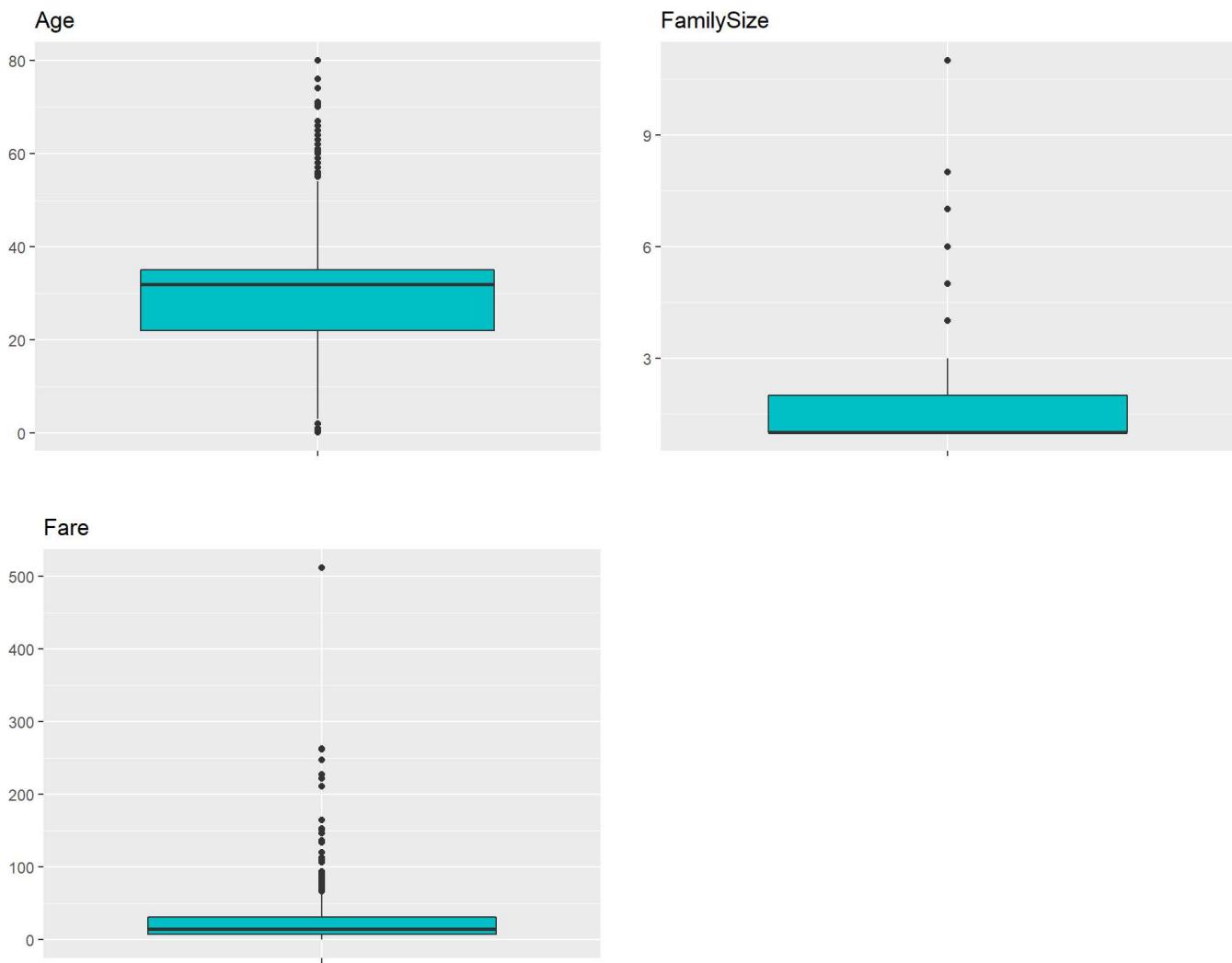
outliers_Age <- ggplot(titanic_data) +
  aes(x = "", y = Age) +
  geom_boxplot(fill = "#00bfc4") +
  ylab("") + labs(x "") +
  ggtitle("Age")

outliers_FamilySize <- ggplot(titanic_data) +
  aes(x = "", y = FamilySize) +
  geom_boxplot(fill = "#00bfc4") +
  ylab("") + labs(x "") +
  ggtitle("FamilySize")

outliers_Fare <- ggplot(titanic_data) +
  aes(x = "", y = Fare) +
  geom_boxplot(fill = "#00bfc4") +
  ylab("") + labs(x "") +
  ggtitle("Fare")

# Mostramos Las gráficas
grid.arrange(outliers_Age, outliers_FamilySize, outliers_Fare, ncol = 2)

```



Las gráficas muestran que existen valores extremos para estas tres variables:

- En el caso de la edad (*Age*), en la gráfica se puede ver como existen valores extremos inferiores y superiores a la franja que va, aproximadamente, de los 20 a los 40 años. Sin embargo, estos valores atípicos entran dentro de lo que podríamos considerar un rango de edad razonable, así que conservaremos estos valores.
- En lo que respecta al tamaño de la familia (*FamilySize*) en la gráfica se observan 6 valores muy alejados del resto, siendo especialmente llamativo el que indica un tamaño de familia de más de 10 miembros. Sin embargo, no parece tratarse de un error, ya que realmente existió una familia a bordo del Titanic cuyos 11 miembros perecieron durante su hundimiento, tal y como se puede ver en este artículo (<https://www.bbc.com/news/uk-england-cambridgeshire-17596264>). En definitiva, todo indica que estos datos son correctos por lo que también vamos a conservarlos.
- En el caso de la variable que almacena el precio del billete (*Fare*), en su gráfica destaca un valor superior a 500, muy por encima del precio del resto de billetes. En este caso vamos a sustituirlo por el precio medio del billete de primera clase mediante el siguiente código:

```
# Seleccionamos Los pasajeros de primera clase que pagaron menos de 500
pasajeros_primeras_titanic <- titanic_data %>% filter(Fare < 500, Pclass == '1')

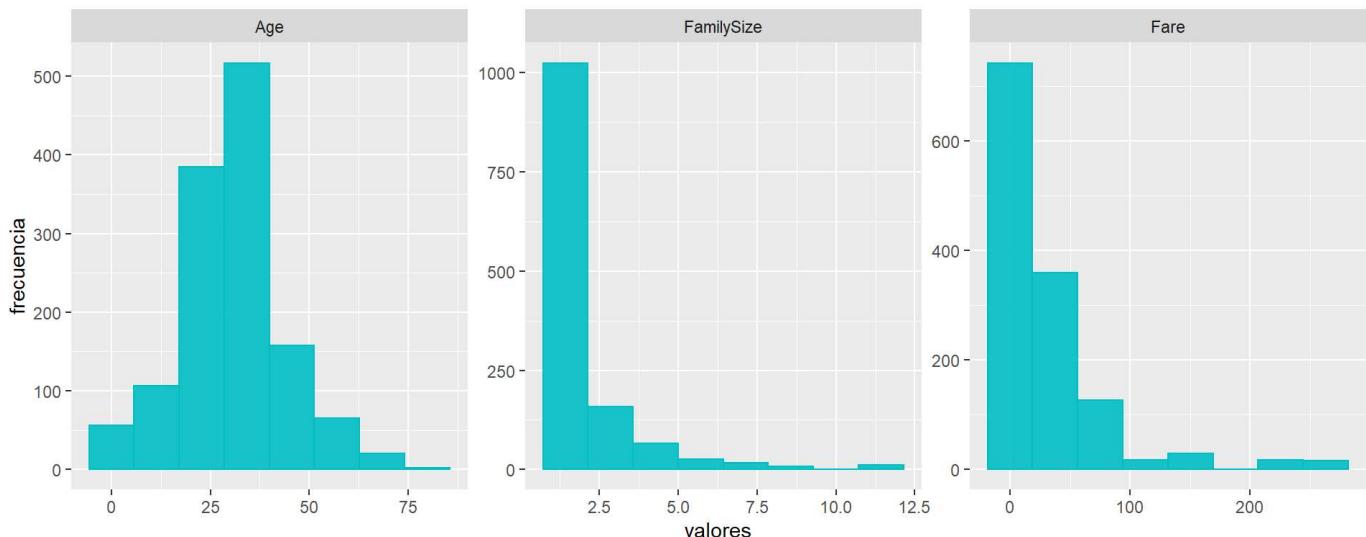
# Calculamos el precio medio del billete pagado por Los pasajeros seleccionados
media_billetes_primeras <- mean(pasajeros_primeras_titanic$Fare, na.rm = TRUE)
media_billetes_primeras
```

```
## [1] 82.18209
```

```
# Sustituimos el precio de los billetes superiores a 500 por el valor calculado
titanic_data$Fare[which(titanic_data$Fare > 500)] <- media_billetes_primera
```

A continuación vamos a mostrar los histogramas de estas variables para tener una idea más clara de la influencia que tienen estos valores atípicos que hemos detectado:

```
# Generamos Los histogramas de Las variables numéricas del dataset
titanic_data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key, scale="free") +
  labs(x = "valores", y = "frecuencia") +
  geom_histogram(fill="#00bfc4", color="#00bfc4", alpha=0.9, bins = 8)
```



Comprobación de la normalidad y homogeneidad de la varianza

En las gráficas anteriores podemos ver cómo las distribuciones de los datos de las variables *FamilySize* y *Fare* están claramente sesgadas a la derecha. En el caso de la variable *Age* la gráfica recuerda vagamente a una “campana de Gauss”. Para salir de dudas vamos a **comprobar la normalidad** de estas tres distribuciones mediante la prueba de *Anderson-Darling*:

```
# Comprobar normalidad de La distribución de Los datos de Age
ad.test(titanic_data$Age)
```

```
##
##  Anderson-Darling normality test
##
## data: titanic_data$Age
## A = 12.945, p-value < 2.2e-16
```

```
# Comprobar normalidad de La distribución de Los datos de FamilySize  
ad.test(titanic_data$FamilySize)
```

```
##  
## Anderson-Darling normality test  
##  
## data: titanic_data$FamilySize  
## A = 171.91, p-value < 2.2e-16
```

```
# Comprobar normalidad de La distribución de Los datos de Fare  
ad.test(titanic_data$Fare)
```

```
##  
## Anderson-Darling normality test  
##  
## data: titanic_data$Fare  
## A = 171.63, p-value < 2.2e-16
```

El *p-value* obtenido en los tres casos es inferior al nivel de significancia 0,05, por lo que podemos rechazar la hipótesis nula y concluir que estas tres distribuciones de datos **no siguen una distribución normal**.

A continuación vamos a comprobar si para estas variables existe homogeneidad en su varianza en cuanto a la supervivencia del pasajero, para ello realizaremos la *prueba de Bartlett*:

```
# Prueba para La variable Age  
bartlett.test(titanic_data$Age, titanic_data$Survived)
```

```
##  
## Bartlett test of homogeneity of variances  
##  
## data: titanic_data$Age and titanic_data$Survived  
## Bartlett's K-squared = 4.4565, df = 1, p-value = 0.03477
```

```
# Prueba para La variable FamilySize  
bartlett.test(titanic_data$FamilySize, titanic_data$Survived)
```

```
##  
## Bartlett test of homogeneity of variances  
##  
## data: titanic_data$FamilySize and titanic_data$Survived  
## Bartlett's K-squared = 57.954, df = 1, p-value = 2.683e-14
```

```
# Prueba para La variable Fare  
bartlett.test(titanic_data$Fare, titanic_data$Survived)
```

```

## 
##  Bartlett test of homogeneity of variances
##
## data: titanic_data$Fare and titanic_data$Survived
## Bartlett's K-squared = 20.794, df = 1, p-value = 5.115e-06

```

En el resultado anterior vemos que los valores *p-value* son inferiores a 0.05, con lo que podemos rechazar la hipótesis nula que establece que las varianzas son iguales, concluyendo que **no existe homogeneidad en la varianza** respecto a la supervivencia del pasajero.

Discretización de la edad de los pasajeros

Para finalizar este apartado, vamos a crear una nueva variable categórica que contendrá los valores discretizados de la variable que contiene la edad de los pasajeros (*Age*), con el objetivo de facilitar el análisis que llevaremos a cabo a continuación.

Esta discretización la vamos a realizar en segmentos de 10 años, finalizando en el intervalo 80-89, ya que la edad máxima registrada para los pasajeros es de 80 años, tal y como demuestra el siguiente código:

```
summary(titanic_data$Age)
```

```

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      0.17    22.00   32.00    30.49   35.00    80.00

```

Las siguientes líneas de código crearán la nueva variable categórica *AgeRange*:

```

# Discretizamos los valores de la variable "Age"

titanic_data["AgeRange"] <- cut(titanic_data$Age, breaks = c(0,9,19,29,39,49,59,69,79,
89), labels = c("0-9","10-19","20-29","30-39","40-49","50-59","60-69","70-79","80-89"
))

```

Una vez finalizada esta etapa de preprocesado, el siguiente fragmento de código almacenará en el disco el *dataset* generado:

```
write.csv(titanic_data,"titanic_data.csv", row.names = FALSE)
```

Análisis de los datos

El primer paso de este análisis será visualizar cómo se distribuyen las distintas variables de este conjunto de datos.

Vamos a comenzar por las variables **variables categóricas**. El siguiente código permite visualizar como están distribuidos sus valores:

```

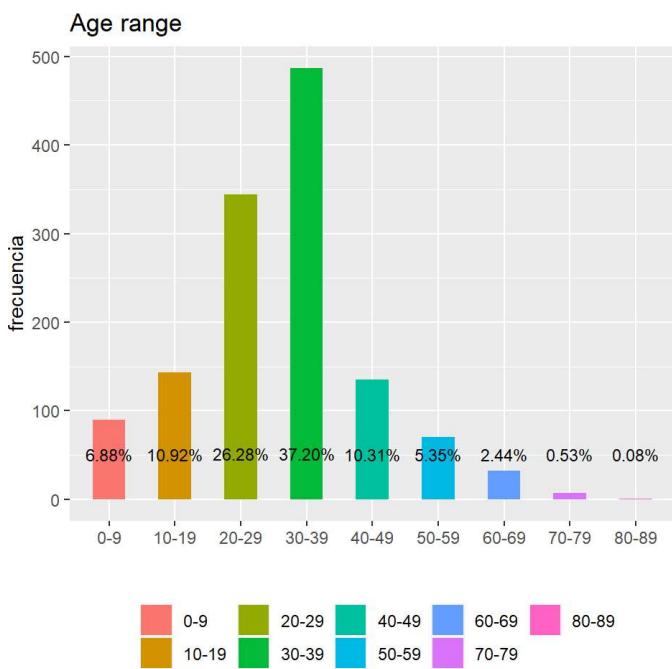
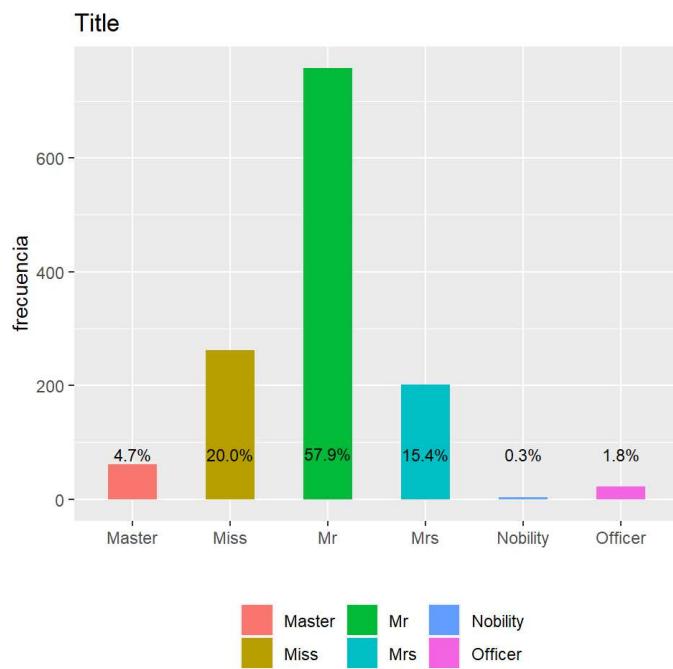
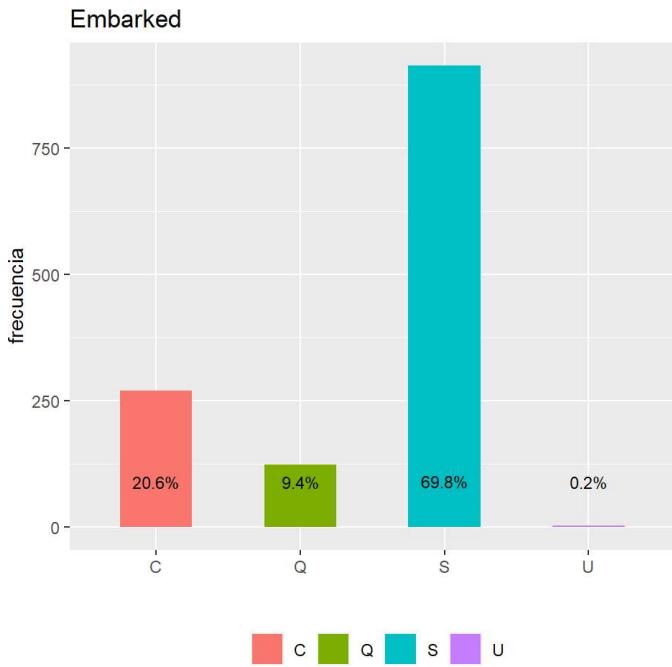
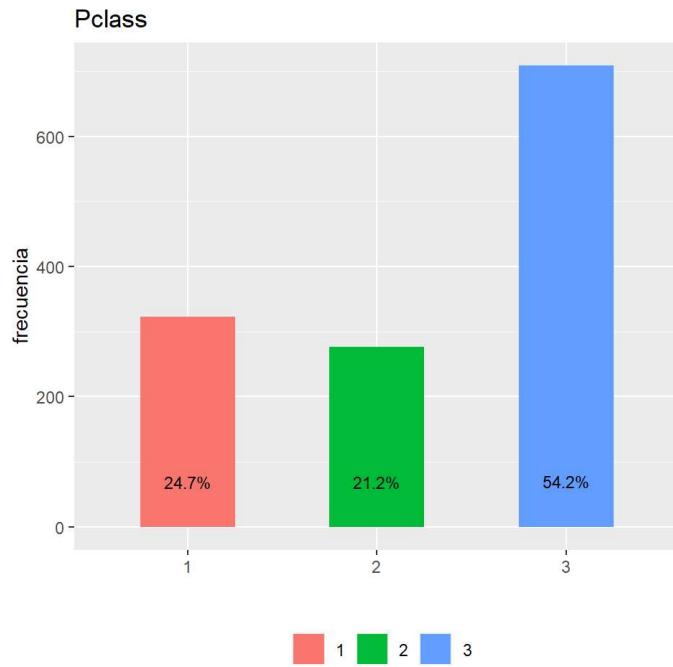
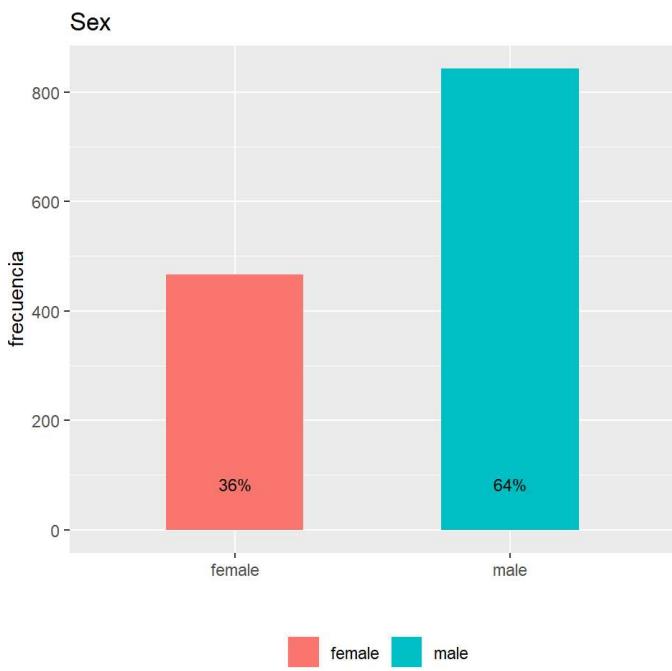
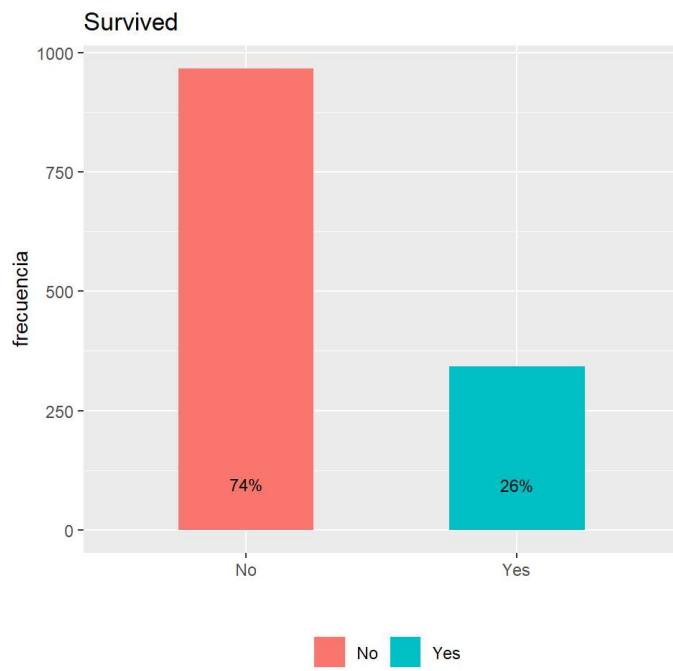
# Definimos una función para generar las gráficas de distribución
mostrarDistribucion <- function(dataset,variable,titulo){

  grafico <- ggplot(dataset,
    aes(x=as.character(variable),fill=as.character(variable))) +
    geom_bar(stat="count", width=0.5, position = 'dodge') +
    geom_text(aes(label=scales::percent(..count..)/sum(..count..)),
              y=(..count..)/sum(..count..)),stat="count",position = position_dodge(width = .9), vjust=-3, size= 3) +
    ylab("frecuencia") + labs(x="",fill "") +
    ggtitle(titulo) +
    theme(legend.position="bottom")
}

# Generamos las gráficas utilizando la función que hemos definido antes
cat_1 <- mostrarDistribucion(titanic_data,titanic_data$Survived,'Survived')
cat_2 <- mostrarDistribucion(titanic_data,titanic_data$Sex,'Sex')
cat_3 <- mostrarDistribucion(titanic_data,titanic_data$Pclass,'Pclass')
cat_4 <- mostrarDistribucion(titanic_data,titanic_data$Embarked,'Embarked')
cat_5 <- mostrarDistribucion(titanic_data,titanic_data>Title,'Title')
cat_6 <- mostrarDistribucion(titanic_data,titanic_data$AgeRange,'Age range')

grid.arrange(cat_1, cat_2, cat_3, cat_4, cat_5, cat_6, ncol = 2)

```

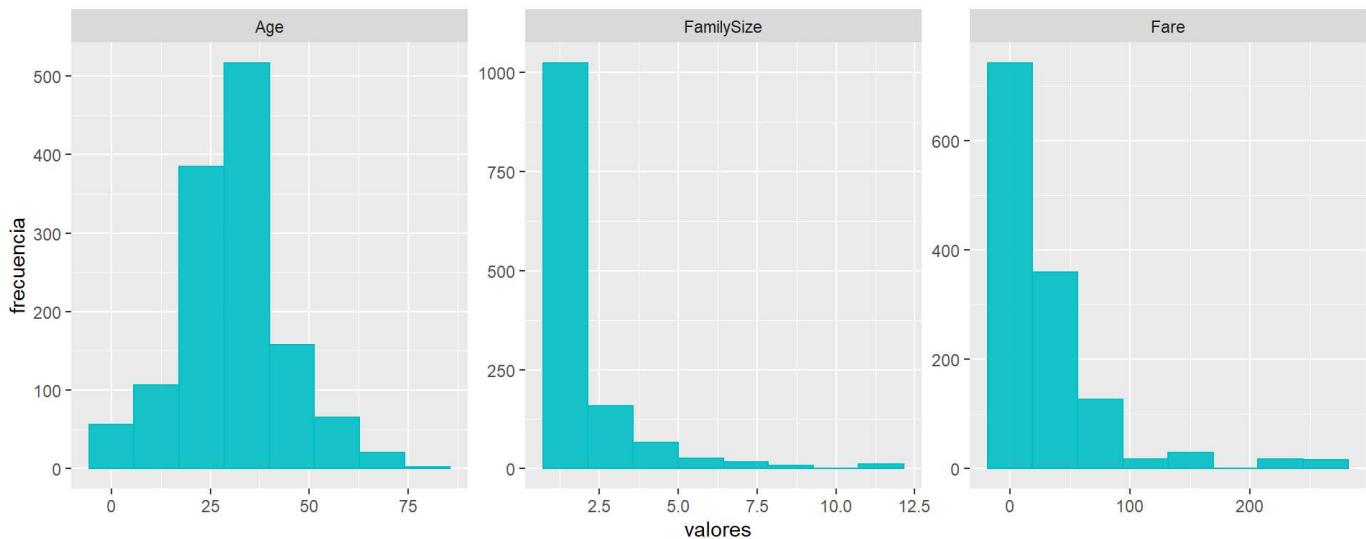


Las gráficas anteriores muestran que:

- La mayoría de los pasajeros perecieron en el naufragio (74%)
- La mayoría de los pasajeros eran hombres (64%)
- Más de la mitad de los pasajeros viajaban en tercera clase.
- La mayoría de los pasajeros embarcaron en Southampton, al sur de Inglaterra.
- La presencia de miembros de la nobleza entre el pasaje era prácticamente inapreciable.
- Más de la mitad de los pasajeros tenía entre 20 y 40 años.
- El porcentaje de niños y adolescentes a bordo era relativamente bajo, apenas el 15% del total de pasajeros.

En lo que respecta a la distribución de las **variables numéricas**, vamos a generar de nuevo sus histogramas para ver qué información pueden ofrecer:

```
# Generamos Los histogramas de Las variables numéricas del dataset
titanic_data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key, scale="free") +
  labs(x = "valores", y = "frecuencia") +
  geom_histogram(fill="#00bfc4", color="#00bfc4", alpha=0.9, bins = 8)
```



A la vista de estas gráficas, algunas de las conclusiones más inmediatas son:

- Se confirma que la edad de la mayoría de los pasajeros estaba entre los 20 y 40 años aproximadamente. Sin embargo, es necesario tener en cuenta que estos datos están influenciados por el reemplazo de más de 200 valores nulos de la variable *Age* que hemos realizado en el apartado anterior.
- La presencia de familias con más de dos miembros embarcados era escasa, es decir, la mayoría de pasajeros no viajaba con niños o parientes.
- La mayoría de billetes vendidos tenían precios relativamente bajos.

A continuación, vamos a generar una serie de gráficas para intentar visualizar las posibles relaciones entre estas variables y la supervivencia de los pasajeros (*Survived*):

```

# Definimos una función para generar las gráficas de las variables categóricas
compararVar <- function(dataset,principal,grupo,titulo){

  grafico <- ggplot(dataset, aes(x={{principal}},group={{grupo}})) +
    geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
    geom_text(aes(label = scales::percent(..prop..),
                  y= ..prop.. ), stat= "count", vjust = -.1, size=3) +
    labs(x="", y = "") +
    facet_grid(~ {{grupo}}) +
    scale_y_continuous(labels = scales::percent) +
    theme(legend.position = "none") +
    ggtitle(titulo)
}

# Generamos las gráficas de las variables categóricas llamando a la función que hemos
# definido antes
rel_1 <- compararVar(titanic_data,titanic_data$Survived,titanic_data$Sex,"Sex vs Survived")
rel_2 <- compararVar(titanic_data,titanic_data$Survived,titanic_data$Pclass,"Pclass vs Survived")
rel_3 <- compararVar(titanic_data,titanic_data$Survived,titanic_data$Embarked,"Embarked vs Survived")
rel_4 <- compararVar(titanic_data,titanic_data$Survived,titanic_data>Title,"Title vs Survived")

# Generamos ahora las gráficas de las variables numéricas

# Supervivencia por edad
rel_5 <- titanic_data %>%
  ggplot(aes(x = Age, fill = Survived)) +
  geom_histogram() +
  labs(title = "Age vs Survived")

# Supervivencia por precio del billete
rel_6 <- titanic_data %>%
  ggplot(aes(x = Fare, fill = Survived)) +
  geom_histogram() +
  labs(title = "Fare vs Survived")

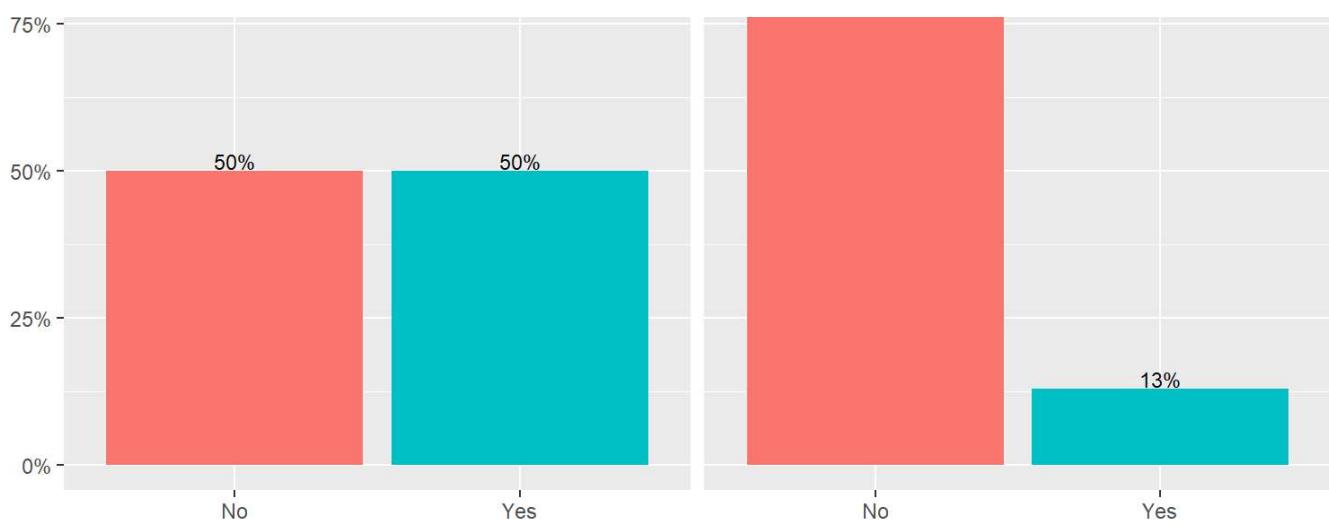
# Supervivencia por tamaño familia
rel_7 <- titanic_data %>%
  ggplot(aes(x = FamilySize, fill = Survived)) +
  geom_histogram() +
  labs(title = "FamilySize vs Survived")

grid.arrange(rel_1, rel_2, rel_3, rel_4, rel_5, rel_6, rel_7, nrow=7)

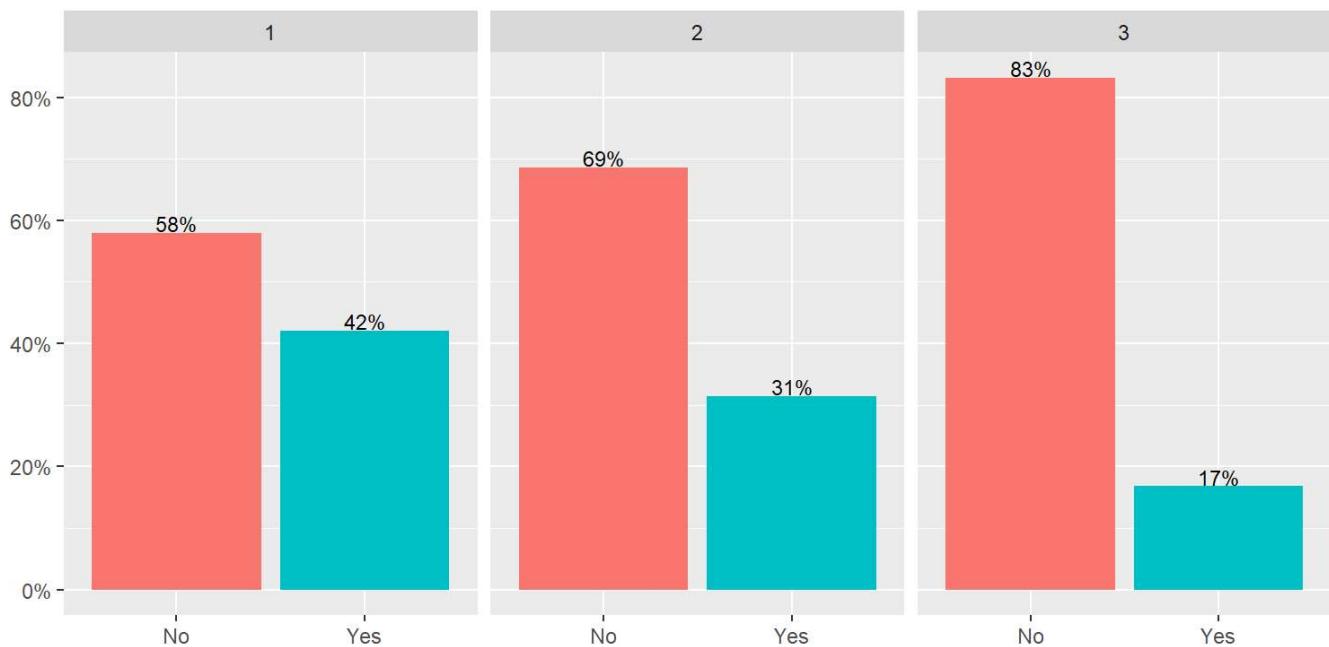
```

Sex vs Survived

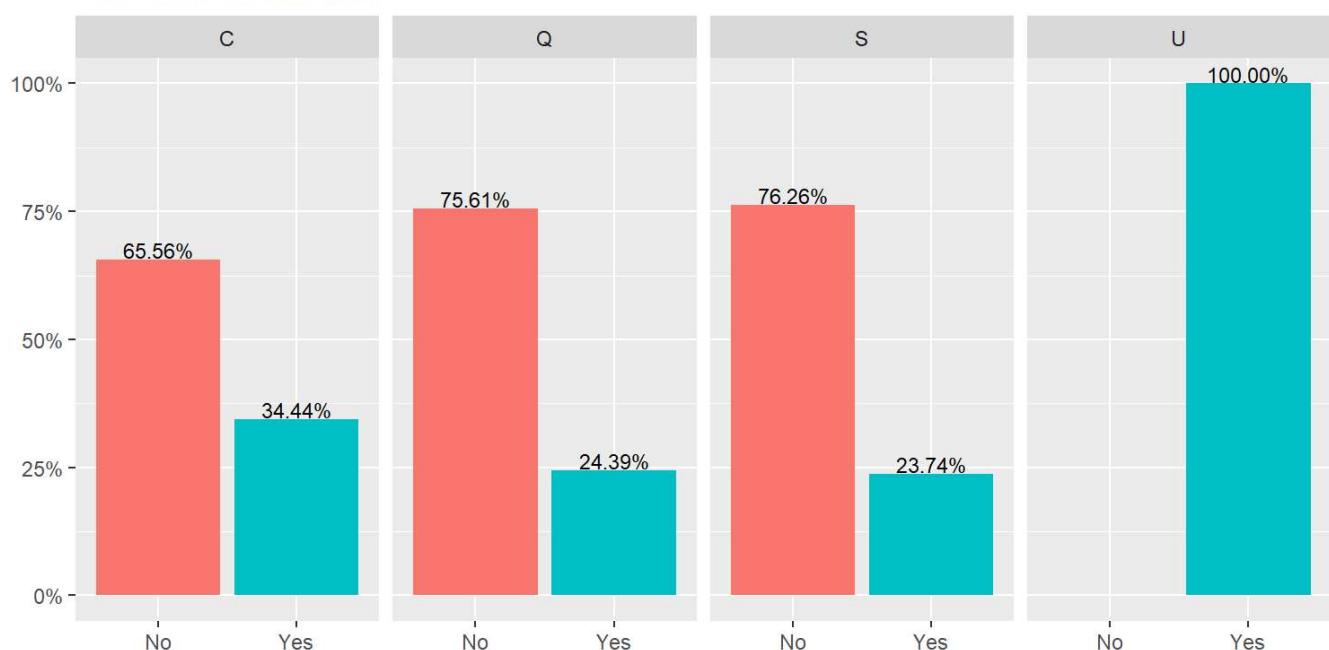




Pclass vs Survived

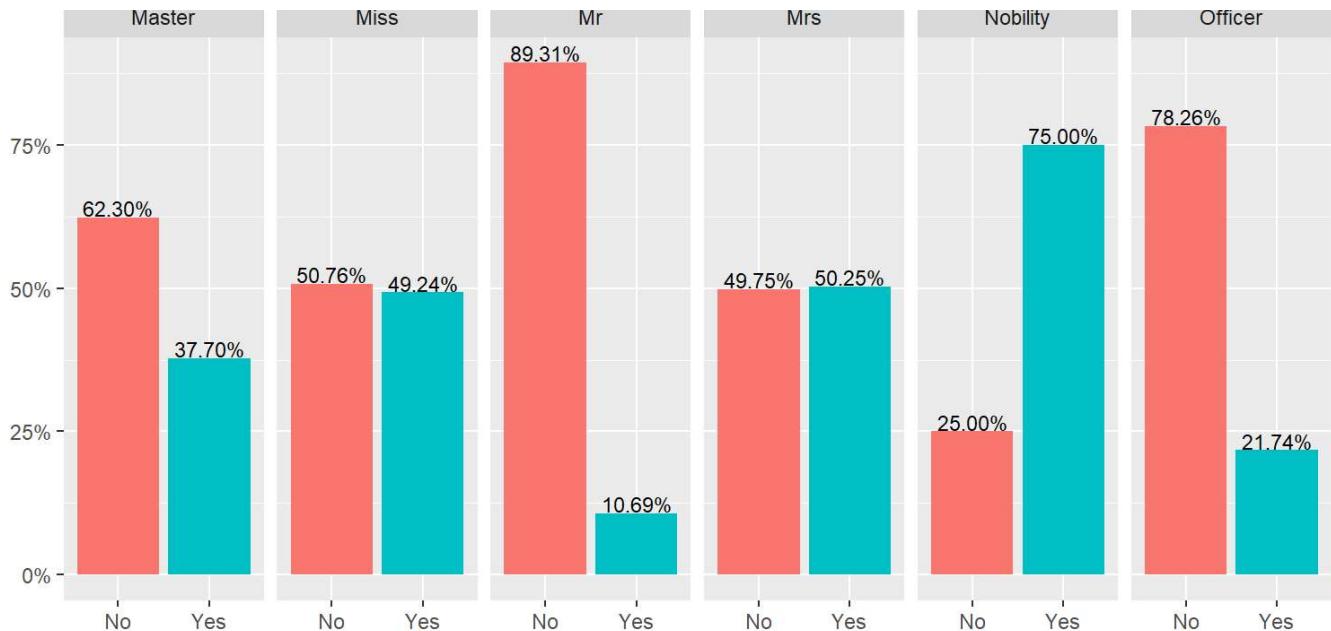


Embarked vs Survived

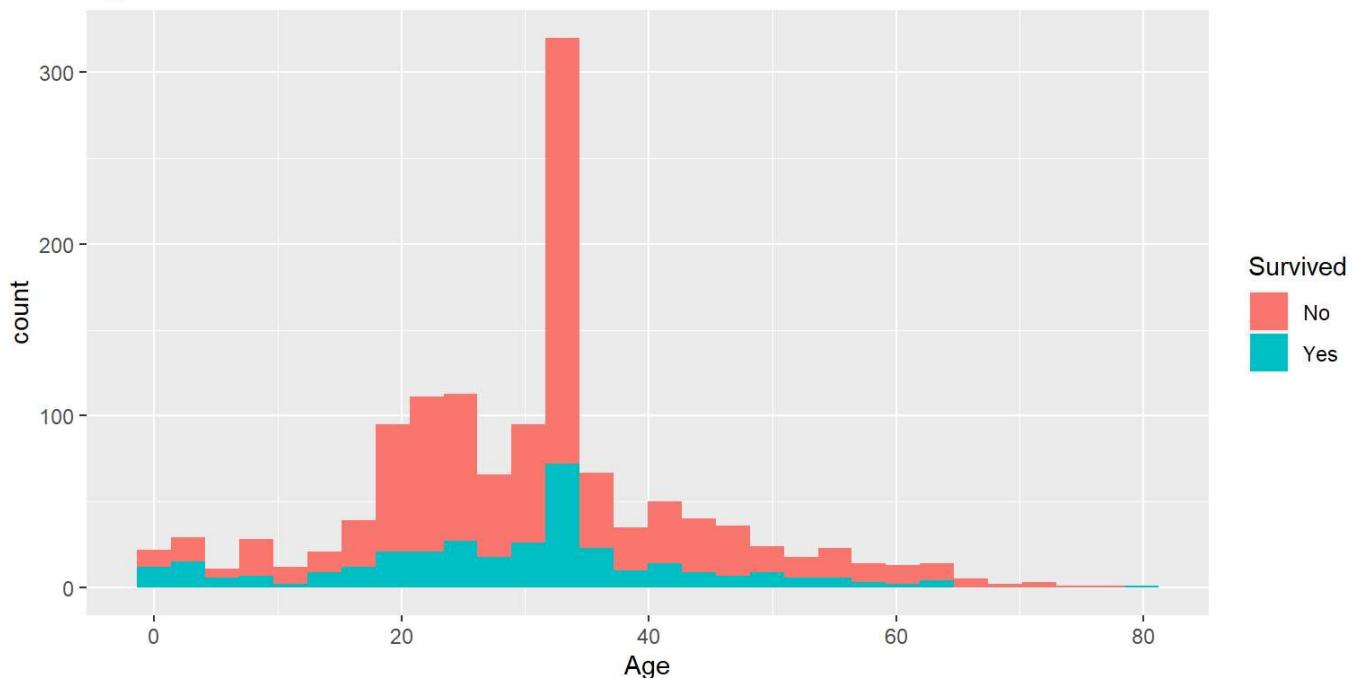


Title vs Survived

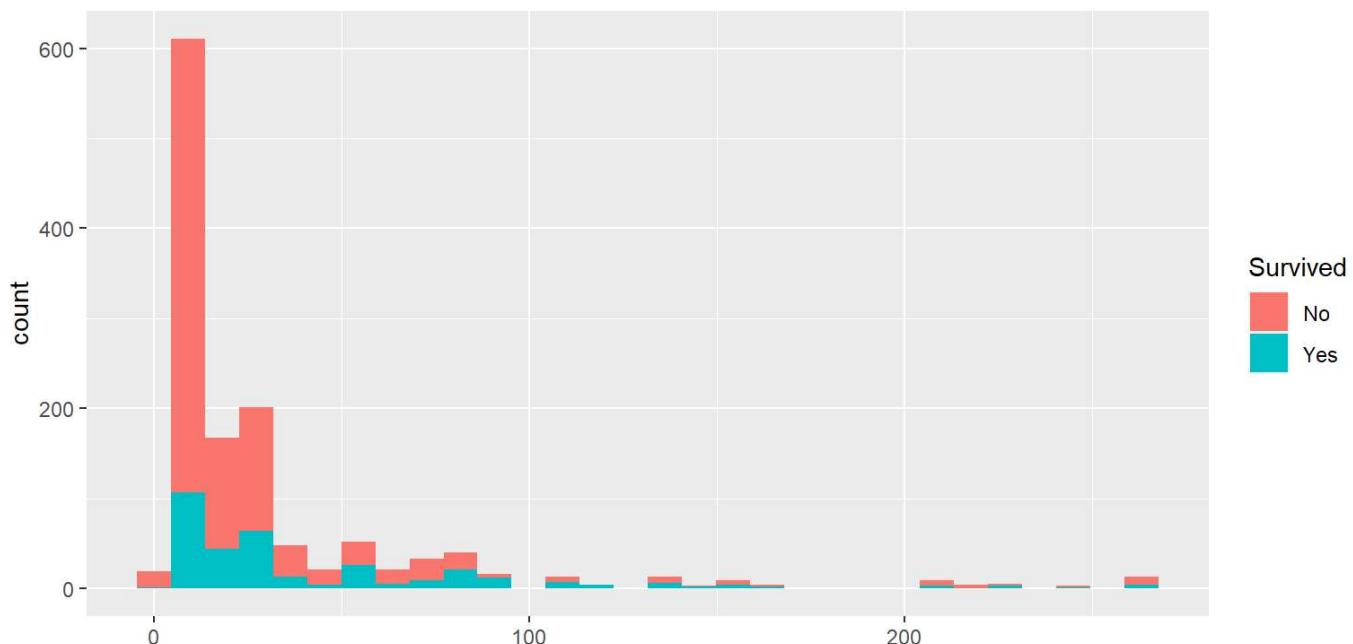


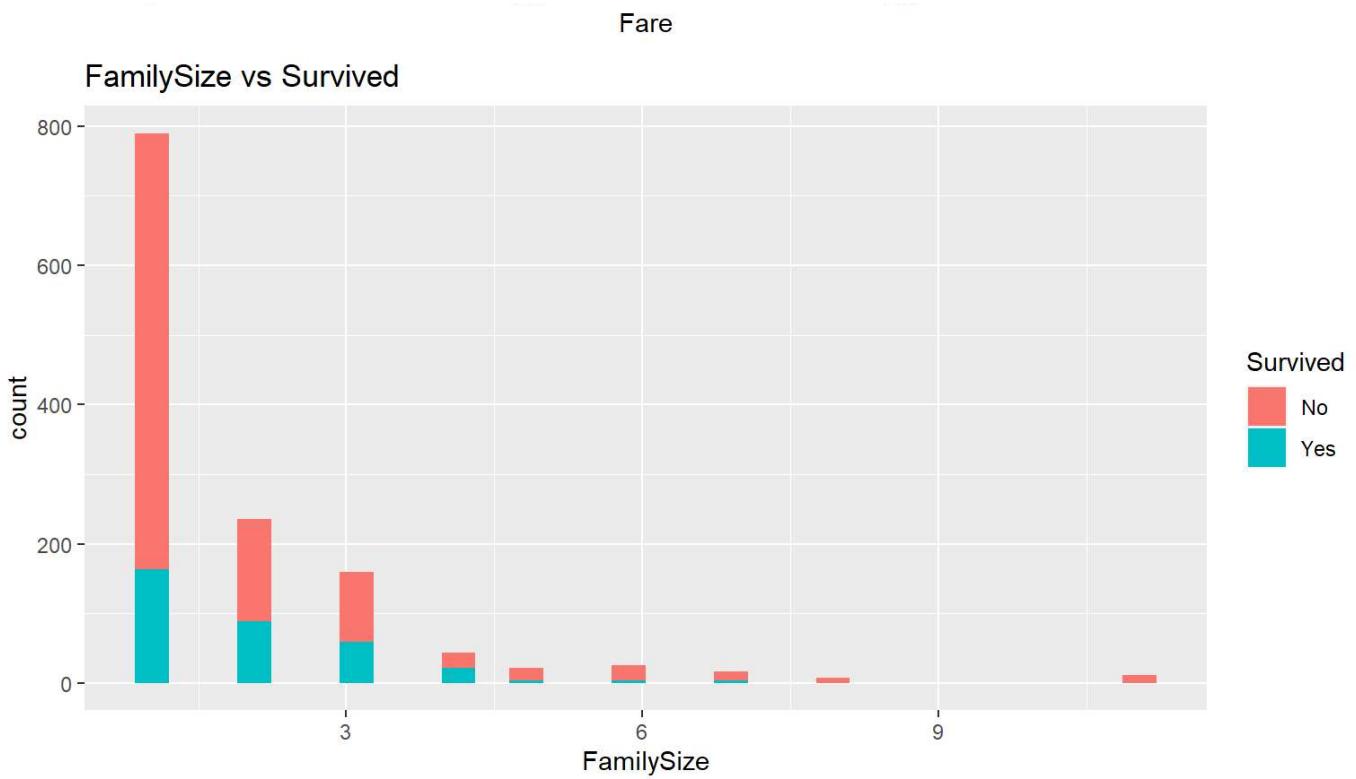


Age vs Survived



Fare vs Survived





El análisis visual de estas gráficas permite sacar algunas conclusiones interesantes:

- La proporción de mujeres que sobrevivieron es considerablemente superior a la de hombres. Sin embargo, **resulta muy extraño** que la proporción de mujeres que sobrevivieron frente a las que no lo consiguieron sea **exactamente del 50%**. Esta circunstancia pone en duda la fiabilidad del segundo conjunto de datos que hemos obtenido de Kaggle para añadir la variable *Survived* (fichero *train_and_test2.csv*) al conjunto de prueba descargado inicialmente (fichero *test.csv*)
- Los pasajeros de tercera clase que consiguieron sobrevivir fueron muy escasos (17%). Sin embargo, si lo hicieron muchos de los que viajaban en primera (42%) y en menor medida los de segunda clase (31%)
- Hubo más supervivientes entre los pasajeros que embarcaron en Francia que los que lo hicieron en un puerto de las Islas Británicas.
- La inmensa mayoría de hombres adultos de clase media o baja perecieron en el hundimiento. Si lograron sobrevivir casi todos los miembros de la nobleza que viajaban en el buque.
- Según se observa en la gráfica, los pasajeros menores de 10 años tuvieron una proporción de supervivencia relativamente alta. Llama la atención como el pasajero más anciano parece que sobrevivió también, tal y como se observa en el valor más a la derecha del eje horizontal de la gráfica.
- También parece que los pasajeros que viajaban con los billetes más caros sobrevivieron en mayor proporción que los que adquirieron los más económicos.
- El aumento del tamaño de la familia embarcada del pasajero parece que reduce sus posibilidades de supervivencia.

Ahora que tenemos cierta intuición de los factores que influyeron en la suerte que corrieron estos pasajeros del Titanic, vamos a comprobar si existe alguna relación de dependencia estadísticamente significativa entre las **variables categóricas** y la que guarda si sobrevivieron o no (*Survived*). Para ello, vamos a generar las tablas de contingencia necesarias para ejecutar una **prueba Chi-cuadrado (χ^2)** a cada una de ellas:

```
# Definimos una función para generar las tablas de contingencia y calcular chi-cuadrado
o
generarTablaContingencia <- function(principal,grupo,titulo){

  tablaCont <- table(principal,grupo)
  chi_cuadrado <- chisq.test(tablaCont)
  titulo <- paste(titulo,". p-value: ", chi_cuadrado$p.value)
  tablaCont <- ggplot() + annotation_custom(tableGrob(tablaCont)) + labs(title = titul
o)

}

t1 <- generarTablaContingencia (titanic_data$Survived,titanic_data$Sex, 'Sex vs Surviva
l')

t2 <- generarTablaContingencia (titanic_data$Survived,titanic_data$Pclass, 'Pclass vs S
urvival')

t3 <- generarTablaContingencia (titanic_data$Survived,titanic_data$Embarked, 'Embarked
 vs Survival')

t4 <- generarTablaContingencia (titanic_data$Survived,titanic_data>Title, 'Title vs Sur
vival')

t5 <- generarTablaContingencia (titanic_data$Survived,titanic_data$AgeRange, 'Age range
 vs Survival')

grid.arrange(t1, t2, t3, t4, t5, nrow=5)
```

Sex vs Survival . p-value: 5.68944265543662e-48

	female	male
No	233	734
Yes	233	109

Pclass vs Survival . p-value: 7.76903815538711e-18

	1	2	3
No	187	190	590
Yes	136	87	119

Embarked vs Survival . p-value: 0.000396392744880634

	C	Q	S	U
No	177	93	697	0
Yes	93	30	217	2

Title vs Survival . p-value: 5.15185122196098e-49

	Master	Miss	Mr	Mrs	Nobility	Officer
No	38	133	677	100	1	18
Yes	23	129	81	101	3	5

Age range vs Survival . p-value: 0.00138354674473047

	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89
No	50	102	267	364	101	50	26	7	0
Yes	40	41	77	123	34	20	6	0	1

En la parte superior de cada tabla se puede ver el valor *p-value* obtenido en la prueba Chi-cuadrado, siendo este **menor que 0,05** para todas las variables, por lo que podemos rechazar la hipótesis nula de que no existe dependencia entre cada una de ellas y *Survived*, por lo que concluimos que sí **existe una dependencia**

significativa.

A continuación vamos a comprobar las posibles dependencias que pueden existir entre las **variables numéricas** de este conjunto de datos, para ello vamos a generar una **matriz de correlación** en la que incluiremos también a la variable *Survived*, "reconvirtiendo sus valores de texto a numeros:

```
# Volvemos a convertir los valores de Survived a números

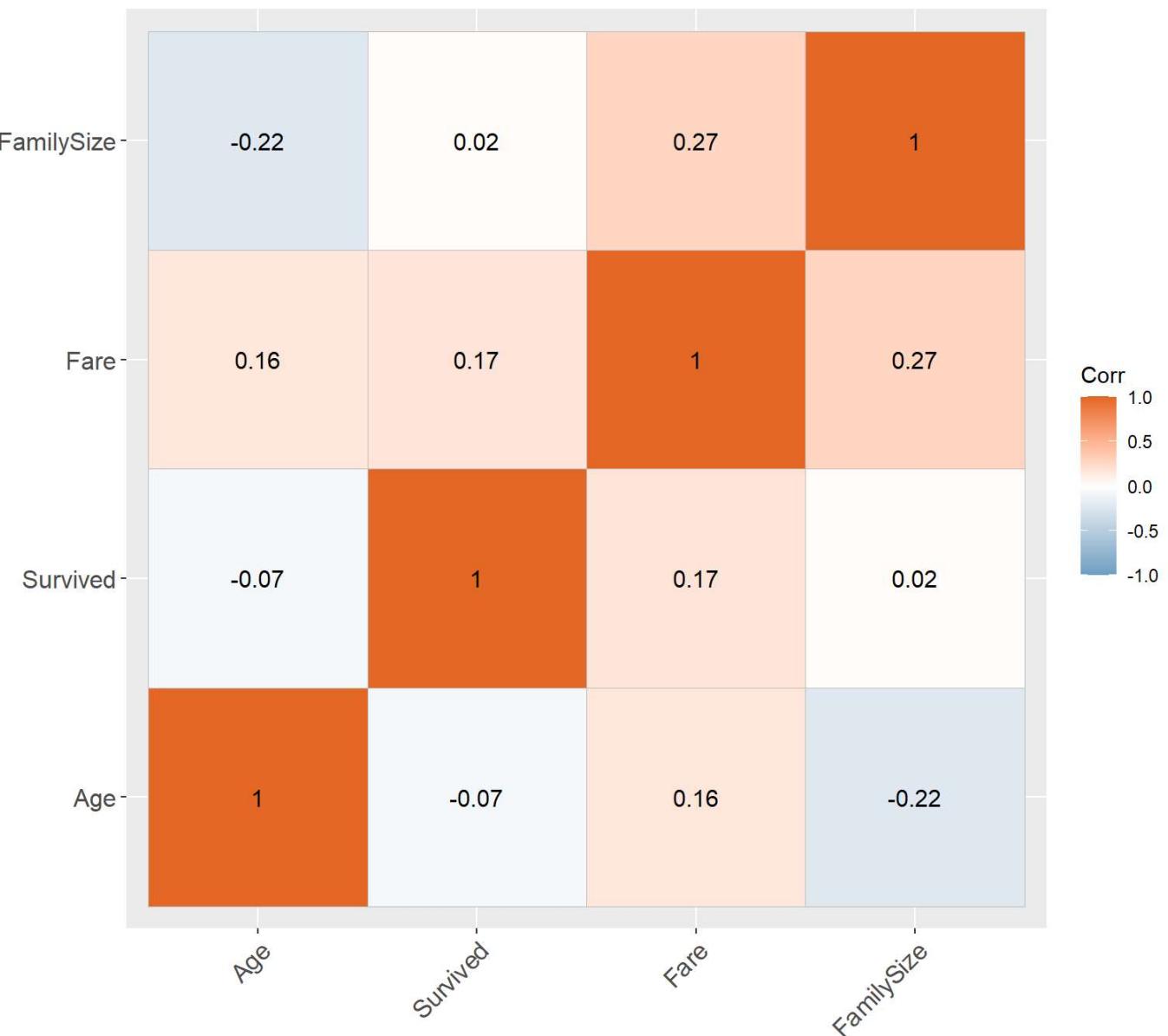
titanic_data$Survived[titanic_data$Survived == "Yes"] <- 1
titanic_data$Survived[titanic_data$Survived == "No"] <- 0

titanic_data$Survived <- as.numeric(titanic_data$Survived)

# Seleccionamos las variables numéricas del dataset
titanic_data_num <- titanic_data %>% select(where(is.numeric))

# Generamos la matriz de correlación
corr <- round(cor(titanic_data_num, method = c("pearson", "kendall", "spearman")), 2)

# Visualizamos la matriz
ggcorrplot(corr, hc.order = TRUE, lab = TRUE,
           ggtheme = ggplot2::theme_gray,
           colors = c("#6D9EC1", "#FFFFFF", "#E46726"))
```



Los coeficientes de la fila de la variable *Survived* muestran como el precio del billete (*Fare*) es la variable numérica con la que tiene una relación más significativa, siendo esta positiva y bastante débil (0,17). La relación con las otras variables es menos evidente, siendo negativa en el caso de la edad del pasajero (*Age*)

En la matriz también llaman la atención los valores de los coeficientes entre las variables *Age* y *FamilySize*, indicando que existe una relación negativa entre ambas, y entre esta última y *Fare*, siendo en este caso el coeficiente con el valor más alto de toda la matriz aunque, al igual que ocurre en las anteriores, es una relación débil.

Modelo de regresión lineal

Para finalizar este análisis vamos a obtener una serie de modelos de regresión lineal que permitan predecir la supervivencia del pasajero utilizando algunas de las variables que hemos analizado en los apartados anteriores.

El siguiente código define un modelo de regresión lineal utilizando únicamente las variables categóricas *Sex* y *Pclass*. Una vez definido se realiza la predicción de la supervivencia de los pasajeros y a continuación se calcula la precisión obtenida.

```
# Generamos el modelo a partir de Sex y Pclass
model1 <- lm(titanic_data$Survived ~ titanic_data$Sex + titanic_data$Pclass, data = titanic_data)

# Predecimos La supervivencia del pasajero mediante el modelo
titanic_data$Survived_pred_m1 <- round(predict(model1, titanic_data))

# Calculamos el porcentaje de observaciones clasificadas correctamente
nrow(titanic_data[titanic_data$Survived_pred_m1 == titanic_data$Survived,]) / nrow(titanic_data)

## [1] 0.7937357
```

El resultado anterior muestra que se obtiene una precisión del 79%. A continuación vamos a generar un nuevo modelo basado en el anterior al que se le añadirá la variable categórica *Title*:

```
# Generamos el modelo a partir de Sex, Pclass y Title
model2 <- lm(titanic_data$Survived ~ titanic_data$Sex + titanic_data$Pclass + titanic_data$title, data = titanic_data)

# Predecimos La supervivencia del pasajero mediante el modelo
titanic_data$Survived_pred_m2 <- round(predict(model2, titanic_data))

# Calculamos el porcentaje de observaciones clasificadas correctamente
nrow(titanic_data[titanic_data$Survived_pred_m2 == titanic_data$Survived,]) / nrow(titanic_data)

## [1] 0.7944996
```

Como se puede ver, la precisión apenas ha aumentado, siendo prácticamente la misma. En el siguiente modelo se añaden al anterior las variables *AgeRange* y *Fare*:

```
# Generamos el modelo a partir de Sex, Pclass, Title, AgeRange y Fare
model3 <- lm(titanic_data$Survived ~ titanic_data$Sex + titanic_data$Pclass + titanic_data$title + titanic_data$AgeRange + titanic_data$Fare, data = titanic_data)

# Predecimos La supervivencia del pasajero mediante el modelo
titanic_data$Survived_pred_m3 <- round(predict(model3, titanic_data))

# Calculamos el porcentaje de observaciones clasificadas correctamente
nrow(titanic_data[titanic_data$Survived_pred_m3 == titanic_data$Survived,]) / nrow(titanic_data)

## [1] 0.7906799
```

En este caso observamos que la precisión desciende ligeramente, por lo que se puede concluir que el mejor modelo de los tres que se han generado es el segundo, en el que se utilizan las variables *Sex*, *Pclass* y *Title* para predecir la supervivencia del pasajero.

Conclusiones

El desarrollo de este caso práctico ha girado en torno a un conjunto de datos que contenía una lista de pasajeros del buque HMS Titanic, hundido trágicamente en abril de 1912 tras colisionar con iceberg. En las distintas etapas de esta práctica se ha llevado a cabo un proceso de carga e integración de datos, tratamiento de valores nulos, desconocidos y extremos del *dataset* y una vez procesados todos los datos se ha llevado a cabo un análisis estadístico y visual de los datos para determinar la influencia que tuvieron las distintas variables registradas en la supervivencia de los pasajeros al hundimiento de este buque, destacando entre todas ellas el sexo del pasajero y la clase en la que viajaba.

Fuentes consultadas

- materiales de la asignatura
- Encyclopedia Titanica
<https://www.encyclopedia-titanica.org/> (<https://www.encyclopedia-titanica.org/>)
- “Replace missing values with R”
<https://www.guru99.com/r-replace-missing-values.html> (<https://www.guru99.com/r-replace-missing-values.html>)
- “Titanic: a deeper look on family size”
<https://www.kaggle.com/lperez/titanic-a-deeper-look-on-family-size>
(<https://www.kaggle.com/lperez/titanic-a-deeper-look-on-family-size>)
- “Titanic Linear Regression”
https://rstudio-pubs-static.s3.amazonaws.com/12743_e0f44945fd1e47a5a8b6d0204264ae9d.html#/
(https://rstudio-pubs-static.s3.amazonaws.com/12743_e0f44945fd1e47a5a8b6d0204264ae9d.html#/)
- “How I Scored in the top 9% of Kaggle’s Titanic Machine Learning Challenge” <https://medium.com/i-like-big-data-and-i-cannot-lie/how-i-scored-in-the-top-9-of-kaggles-titanic-machine-learning-challenge-243b5f45c8e9> (<https://medium.com/i-like-big-data-and-i-cannot-lie/how-i-scored-in-the-top-9-of-kaggles-titanic-machine-learning-challenge-243b5f45c8e9>)
- “Titanic simple rf with name and age features”
<https://www.kaggle.com/ianwells/titanic-simple-rf-with-name-and-age-features>
(<https://www.kaggle.com/ianwells/titanic-simple-rf-with-name-and-age-features>)