Limpieza y Analisis datos: Caso Hundimiento Titanic.

David de Vega Martin

Diciembre 2021

Contents

Actuaciones previas.

Descripcion del conjunto de datos.

El conjunto de datos seleccionado contiene parte de la lista de pasajeros que viajaban en el Titanic en su primer y unico viaje entre Southampton y Nueva York.

En la noche del 14 a 15 de abril de 1912 se hunde en las proximidades de Terranova, tras colisionar con un iceberg.

Como resultado 1.496 personas mueren como consecuencia de lesiones producidas al desalojar el barco, por la posterior hipotermia y por ultimo por ahogamiento.

El dataset que nos ocupa contiene la lista de pasajeros recopilada por Michael A.Findlay y publicada por la Encyclopedia Titanica (http://encyclopedia-titanica.org).

Dos son las razones para elegir este dataset:

Por una parte su popularidad y lo didactico que resulta para tecnicas de Machine Learning que impliquen establecer las probabilidades de supervivencia en base a las variables que presenta el dataset.

Finalmente hay una norma no escrita de que todo buen data-scientist, que se precie, lo ha estudiado alguna vez en su vida y acercandome al ecuador del master era un buen momento para hacerlo.

Carga librerias

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('tidyverse')) install.packages('tidyverse');library('tidyverse')
if (!require('grid')) install.packages('grid');library('grid')
if (!require('lattice')) install.packages('lattice');library('lattice')
if (!require('scales')) install.packages('scales');library('scales')
if (!require('knitr')) install.packages('knitr');library('knitr')
if (!require('gplots')) install.packages('gplots');library('gplots')
if (!require('gmodels')) install.packages('gmodels');library ('gmodels')
if (!require('data.table')) install.packages('data.table');library('data.table')
if (!require('readxl')) install.packages('readxl');library('readxl')
if (!require('corrplot')) install.packages('corrplot');library('corrplot')
if (!require('pander')) install.packages('pander');library('pander')
if(!require('ggpubr')){install.packages('ggpubr');library('ggpubr')}
```

```
if(!require('gridExtra')){install.packages('gridExtra');library('gridExtra')}
if(!require('CGPfunctions')){install.packages('CGPfunctions');library('CGPfunctions')}
if(!require('ggcorrplot')) {install.packages('ggcorrplot'); library('ggcorrplot')}
if (!require('tidyr')) {install.packages('tidyr'); library('tidyr')}
if (!require('purrr')) {install.packages('purrr'); library('purrr')}
if (!require('nortest')) {install.packages('nortest'); library('nortest')}
```

Carga dataset.

```
titanic_test <- read.csv('titanic_test.csv', header=TRUE, stringsAsFactors=FALSE)
titanic_train <- read.csv('titanic_train.csv', header=TRUE, stringsAsFactors=FALSE)
titanic_baseline <- read.csv('gender_baseline.csv', header=TRUE, stringsAsFactors=FALSE)
```

```
str(titanic_train)
```

```
## 'data.frame':    850 obs. of  15 variables:
##  $ passenger_id: int  1216 699 1267 449 576 1083 898 560 1079 908 ...
##  $ pclass      : int  3 3 3 2 2 3 3 2 3 3 ...
##  $ name        : chr  "Smyth, Miss. Julia" "Cacic, Mr. Luka" "Van Impe, Mrs. Jean Baptiste (Rosalie
##  $ sex         : chr  "female" "male" "female" "female" ...
##  $ age         : num  NA 38 30 54 40 28 19 30 22 21 ...
##  $ sibsp       : int  0 0 1 1 0 0 0 0 0 1 ...
##  $ parch       : int  0 0 1 3 0 0 0 0 0 0 ...
##  $ ticket      : chr  "335432" "315089" "345773" "29105" ...
##  $ fare        : num  7.73 8.66 24.15 23 13 ...
##  $ cabin       : chr  "" "" "" "" ...
##  $ embarked    : chr  "Q" "S" "S" "S" ...
##  $ boat        : chr  "13" "" "" "4" ...
##  $ body        : num  NA NA NA NA NA 173 NA NA NA NA ...
##  $ home.dest   : chr  "" "Croatia" "" "Cornwall / Akron, OH" ...
##  $ survived    : int  1 0 0 1 0 0 0 1 1 0 ...
```

```
str(titanic_test)
```

```
## 'data.frame':    459 obs. of  14 variables:
##  $ passenger_id: int  295 1150 89 1063 1020 747 368 1047 569 232 ...
##  $ pclass      : int  1 3 1 3 3 3 2 3 2 1 ...
##  $ name        : chr  "Thayer, Mr. John Borland Jr" "Risien, Mr. Samuel Beard" "Davidson, Mr. Thornto
##  $ sex         : chr  "male" "male" "male" "male" ...
##  $ age         : num  17 NA 31 41 21 ...
##  $ sibsp       : int  0 0 1 0 0 0 1 0 0 0 ...
##  $ parch       : int  2 0 0 0 0 2 0 0 0 0 ...
##  $ ticket      : chr  "17421" "364498" "F.C. 12750" "SOTON/O2 3101272" ...
##  $ fare        : num  110.88 14.5 52 7.12 7.9 ...
##  $ cabin       : chr  "C70" "" "B71" "" ...
##  $ embarked    : chr  "C" "S" "S" "S" ...
##  $ boat        : chr  "B" "" "" "" ...
##  $ body        : num  NA NA NA NA NA NA 17 NA NA 207 ...
##  $ home.dest   : chr  "Haverford, PA" "" "Montreal, PQ" "Finland Sudbury, ON" ...
```

```
str(titanic_baseline)
```

```
## 'data.frame':    459 obs. of  2 variables:
##  $ passenger_id: int  295 1150 89 1063 1020 747 368 1047 569 232 ...
##  $ survived    : int  0 0 0 0 0 0 0 1 0 0 ...
```

Integracion de los datos.

En el dataset train tenemos 850 observaciones con 15 variables, en el test tenemos 459 observaciones con 14 variables, falta la variable survived. Algo logico puesto que este dataset se utiliza para predecir la probabilidad de supervivencia de los viajeros.

En el dataset baseline tenemos la lista completa de los viajeros y la informacion de si sobrevivieron o no. Por esta razon, la primera tarea es reconstruir esa informacion y luego unir ambos dataset en uno solo sobre el que procederemos a realizar nuestro (EDA) Analisis Exploratorio.

```
titanic_test$survived <-NA # Creamos la variable vacia en el dataset.

# Cargamos la variable con los valores de baseline.
# Podriamos hacer un join como metodo alternativo.
titanic_test$survived <-
titanic_baseline$survived[titanic_baseline$passenger_id %in% titanic_test$passenger_id]
```

Unimos los datasets y comprobamos que no existen los duplicados.

```
titanic = rbind(titanic_train,titanic_test)
anyDuplicated(titanic) # Comprobamos la existencia de duplicados.
```

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

```
str(titanic)
```

```
## 'data.frame':    1309 obs. of  15 variables:
##  $ passenger_id: int  1216 699 1267 449 576 1083 898 560 1079 908 ...
##  $ pclass      : int  3 3 3 2 2 3 3 2 3 3 ...
##  $ name        : chr  "Smyth, Miss. Julia" "Cacic, Mr. Luka" "Van Impe, Mrs. Jean Baptiste (Rosalie I
##  $ sex         : chr  "female" "male" "female" "female" ...
##  $ age         : num  NA 38 30 54 40 28 19 30 22 21 ...
##  $ sibsp       : int  0 0 1 1 0 0 0 0 0 1 ...
##  $ parch       : int  0 0 1 3 0 0 0 0 0 0 ...
##  $ ticket      : chr  "335432" "315089" "345773" "29105" ...
##  $ fare        : num  7.73 8.66 24.15 23 13 ...
##  $ cabin       : chr  "" "" "" "" ...
##  $ embarked    : chr  "Q" "S" "S" "S" ...
##  $ boat        : chr  "13" "" "" "4" ...
##  $ body        : num  NA NA NA NA NA 173 NA NA NA NA ...
##  $ home.dest   : chr  "" "Croatia" "" "Cornwall / Akron, OH" ...
##  $ survived    : int  1 0 0 1 0 0 0 1 1 0 ...
```

Tenemos un total de 1309 observaciones con 15 variables que obedecen a la siguiente descripcion:

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

- *PassengerId*: Identificador numérico del pasajero.
- *Survived*: Variable categórica.El pasajero, si sobrevivió valor "1", valor "0" si fallecio. Se trata de la variable que clasifica las observaciones de este *dataset*.
- *Pclass*: Variable categórica. Los valores son clases; "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. Nombre, apellidos y tratamiento del viajero (Mr, Miss, etc.)
- *Sex*: Variable categórica. Recoge el sexo del pasajero. Sus valores son "*male*" para los hombres y "*female*" para las mujeres.
- *Age*: Variable numerica. Edad del pasajero en años. Contiene decimales cuando se trata de menores de un año o es un valor estimado.
- *SibSp*: Variable numérica. El número de familiares adultos del pasajero que se encontraban a bordo (hermanos, cónyuge, etc.).
- *Parch*: Variable numérica. En el caso de los niños contiene el número de progenitores que se encontraban a bordo. En los adultos el número de hijos que viajaban con él.
- *Ticket*: Variable de texto.Numero del billete del pasajero.
- *Fare*: Valor numérico decimal. 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. Los valores desconocidos los recogeremos como "U" de Unknown.

Tratamiento de NAS y valores nulos:

```
colSums(is.na(titanic))
```

```
## passenger_id       pclass         name          sex          age        sibsp
##            0            0            0            0          263            0
##        parch       ticket         fare        cabin     embarked         boat
##            0            0            1            0            0            0
##         body    home.dest     survived
##         1188            0            0
```

```
colSums(titanic =="")
```

```
## passenger_id       pclass         name          sex          age        sibsp
##            0            0            0            0           NA            0
##        parch       ticket         fare        cabin     embarked         boat
##            0            0           NA         1014            2          823
##         body    home.dest     survived
##           NA          564            0
```

Tenemos campos con NA o cadenas vacias para age, fare, cabin, embarked, boat, body y home.dest.

Para age, fare vamos a realizar imputaciones de informacion a partir de los datos del resto del dataset.

Para las variables body, boat y home.dest, falta demasiada informacion que no podemos reconstruir, por lo que no nos proporcionan informacion alguna.

Body hace refencia al numero de los cuerpos recuperados. Como la gran mayoria no fueron recuperados, o de serlo fueron enterrados en el mar, no puede hacerse mucho mas trabajo que comparar el numero de cuerpos recuperados y analizar que sexo, edad y clase pertenecian.

Boat hace referencia a en que bote fueron rescatados los pasajeros supervivientes. Esta variable puede servirnos para obtener metricas de ocupacion de cada bote. Asimismo puede obtenerse que personas fueron rescatadas por un barco determinado ya que se conoce dicho dato.

En cuanto home.dest hace referencia a cual era el destino final de cada pasajero. La gran mayoria de esa informacion es desconocida y no tiene utilidad alguna para el estudio que vamos a desarrollar.

Cabin recoge el numero de camarote que ocupaba cada viajero. Dicha informacion para los viajeros de tercera clase es directamente inexistente puesto que viajaban en camarotes comunales para dicha clase. En cuanto a primera y segunda, el listado esta muy incompleto. Se podria, a partir de los datos de precio pagado por billete intentar establecer que camarotes podian ocupar sus viajeros. Ni siquiera en primera y segunda clase, el precio de los billetes en funcion del punto de partida era unico, ya que dependia de las comodidades extras que tenian incluso dentro de la misma clase o de los servicios a los que tenian derecho.

Asi tenemos constancia de que existian suites de mayor tamaño para que los mas pudientes pudiesen alojar no solo a su familia si no a todos sus empleados domesticos. Tambien hay constancia de a existencia de terrazas privadas y determinadas atenciones por parte de los trabajadores del barco. Esto hace que sea muy laborioso para los objetivos de este estudio analizarlas.

Como consecuencia de todo lo expuesto anteriormente decidimos eliminar todas estas variables de nuestro dataset.

```
titanic$body <- NULL
titanic$boat <- NULL
titanic$home.dest <-NULL
titanic$cabin <-NULL
titanic$passenger_id <-NULL
```

Limpieza de datos.

Para los registros vacios de embarked creamos la categoria “U” Unknown.

```
titanic$embarked[is.na (titanic$embarked) | titanic$embarked == ""] <- "Unkown"
```

Tras esto nos quedan por completar age y fare.

Procedemos a buscar los registros NA de fare. Observamos que son viajeros de tercera clase y sustituimos los valores por la media del precio de tercera clase.

```
# Identificamos los valores fare = NA
titanic$passenger_id[is.na(titanic$fare)]
```

```
## NULL
```

```
titanic$passenger_id[1225]
```

```
## NULL
```

```
# Calculamos la media y sustituimos.
mean_third <- round(mean(titanic$fare[titanic$pclass == "3"], na.rm =TRUE),2)
mean_third
```

```
## [1] 13.3
```

```
titanic$fare[is.na(titanic$fare == "NA")]  <- mean_third
```

Para la edad vamos a extraer del nombre de los viajeros el trato que reciben Master, Miss etc para intentar determinar grupos de edad, en funcion de este tratamiento e imputar los datos faltantes en funcion de dicho tratamiento para los valores que faltan.

Este detalle que parece nimio, modifica bastante los resultados posteriores. Tenemos una escala muy amplia, con viajeros que no llegaban al año, mientras que tenemos otros que alcanzan los 80 años.

Como veremos posteriormente la edad media del pasaje no era elevada, pero en determinados grupos de edad si es palpable la diferencia que supone una imputacion al tramo de edad mas cercano.

En el caso de hombres distingumos de modo generico dos categorias Mister como tratamiento por defecto y Officer que recoge a militares, medicos y clerigos que viajaban entre el pasaje.

En el caso de las mujeres distinguimos entre Mrs para las casadas y Miss para las solteras.

Por reduccion nos queda Master como denominacion para los menores de edad, que ademas hemos comprobado que era la que efectivamente se empleaba en aquella epoca.

Asi mediante una expresion regular extraemos el tratamiento que esta contenido en el nombre completo los agrupamos y posteriormente se procede a discretizar para reducir el numero de categorias inicial conforme a lo expuesto anteriromente.

```
title <-  gsub("^.*, (.*?)\\..*$", "\\1", titanic$name)
titanic$title <- as.factor(title)
table(title)
```

```
## title
##         Capt          Col          Don         Dona           Dr     Jonkheer
##            1            4            1            1            8            1
##         Lady        Major       Master         Miss         Mlle          Mme
##            1            2           61          260            2            1
##           Mr          Mrs           Ms          Rev          Sir the Countess
##          757          197            2            8            1            1
```

Tenemos demasiados niveles los vamos a reducir:

```
levels(titanic$title) <- list(

  # Agrupamos a militares, medicos y clerigos como Officer:
  Officer = c('Capt','Col','Major','Dr','Rev'),

  # Agrupamos los titulos nobiliarios como "Nobility":
  Nobility = c('Jonkheer','Sir','the Countess','Lady'),

  # Agrupamos a los restantes hombres com 'Mr':
  Mr = c('Mr','Don'),

  # Agrupamos a todas las mujeres casadas:
 Mrs =  c('Mrs','Mme','Ms','Dona'),

  # El resto son las mujeres solteras:
 Miss = c('Miss','Mlle'),
```

```
 # Por ultimo se solia referir a los niños de la epoca como master
 Master = c('Master'))
unique(titanic$title)
```

```
## [1] Miss     Mr       Mrs      Master   Officer  Nobility
## Levels: Officer Nobility Mr Mrs Miss Master
```

Calculamos la media de la edad de cada grupo y la aplicamos:

```
mean_age <-aggregate(x = titanic$age,
                     by = list(titanic$title),
                     FUN = mean,na.action = na.omit)

mean_mr <- round(mean(titanic$age[titanic$title == "Mr"], na.rm =TRUE),2)
mean_mrs <- round(mean(titanic$age[titanic$title == "Mrs"], na.rm =TRUE),2)
mean_miss <- round(mean(titanic$age[titanic$title == "Miss"], na.rm =TRUE),2)
mean_officer <- round(mean(titanic$age[titanic$title == "Officer"], na.rm =TRUE),2)
mean_master <- round(mean(titanic$age[titanic$title == "Master"], na.rm =TRUE),2)
mean_nobility <- round(mean(titanic$age[titanic$title == "Nobility"], na.rm =TRUE),2)

titanic$age[is.na(titanic$title== "Mr" & titanic$age == "NA")]  <- mean_mr
titanic$age[is.na(titanic$title== "Mrs" & titanic$age == "NA")]  <- mean_mrs
titanic$age[is.na(titanic$title== "Miss" & titanic$age == "NA")]  <- mean_miss
titanic$age[is.na(titanic$title== "Officer" & titanic$age == "NA")]  <- mean_officer
titanic$age[is.na(titanic$title== "Master" & titanic$age == "NA")]  <- mean_master
titanic$age[is.na(titanic$title== "Nobility" & titanic$age == "NA")]  <- mean_nobility
```

Tras eliminar, previa justificacion de motivos, unas variables y pre procesar otras hemos concluido esta fase no si antes comprobar que nuestro dataset no tiene NA ni cadenas vacias.

```
colSums(is.na(titanic))
```

```
##   pclass     name      sex      age    sibsp    parch   ticket     fare
##        0        0        0        0        0        0        0        0
## embarked survived    title
##        0        0        0
```

```
colSums(titanic =="")
```

```
##   pclass     name      sex      age    sibsp    parch   ticket     fare
##        0        0        0        0        0        0        0        0
## embarked survived    title
##        0        0        0
```

Tras todo este proceso tenemos nuestro dataset inicial,listo para empezar a trabajar con el.

En primer lugar vamos a obtener algunas variables derivadas de otras que por si solas no nos proporcionan una informacion lo suficientemente clara, tal es el caso de las variables SibSp y Parch.

Recordamos que SibSp es el numero de familiares adultos de un pasajero ya sea conyuge, hermanos etc. Por otra parte, SibSp para el caso de los niños recoge el numero de progenitores que viajaban con ellos a bordo.

Procederemos a unificarlas para crear unidades familiares:

Primero unimos ambas variables,para crear unidades familiares posteriormente procedemos por una parte a crear una variable que nos permita hacer calculos numericos y por otra parte discretizaremos en familias que tengan unos determinados tamaños.

Distinguimos entre solteros singles, familias de hasta 4 miembros small, de 5 miembros o mas, big.

```
titanic$familyT <- titanic$sibsp + titanic$parch + 1 # Unimos las variables
titanic$family[titanic$familyT == 1] <- 'Single' # Solteros.
titanic$family[titanic$familyT >= 2  &
               titanic$familyT <  5 ] <- 'Small' # Small. Menos de 5 miembros
titanic$family[titanic$familyT >= 5 ] <- 'Big'   # Big. Mas de 5 miembros.
```

Efectuado esto nos quedamos con las variables de familia una categorica y otra numerica para el EDA. Podemos eliminar sibsp y parch.

```
titanic$sibsp <- NULL
titanic$parch <- NULL
```

Para facilitar la lectura de la variable survived pasamos a yes y no. Transformamos pclass en factor

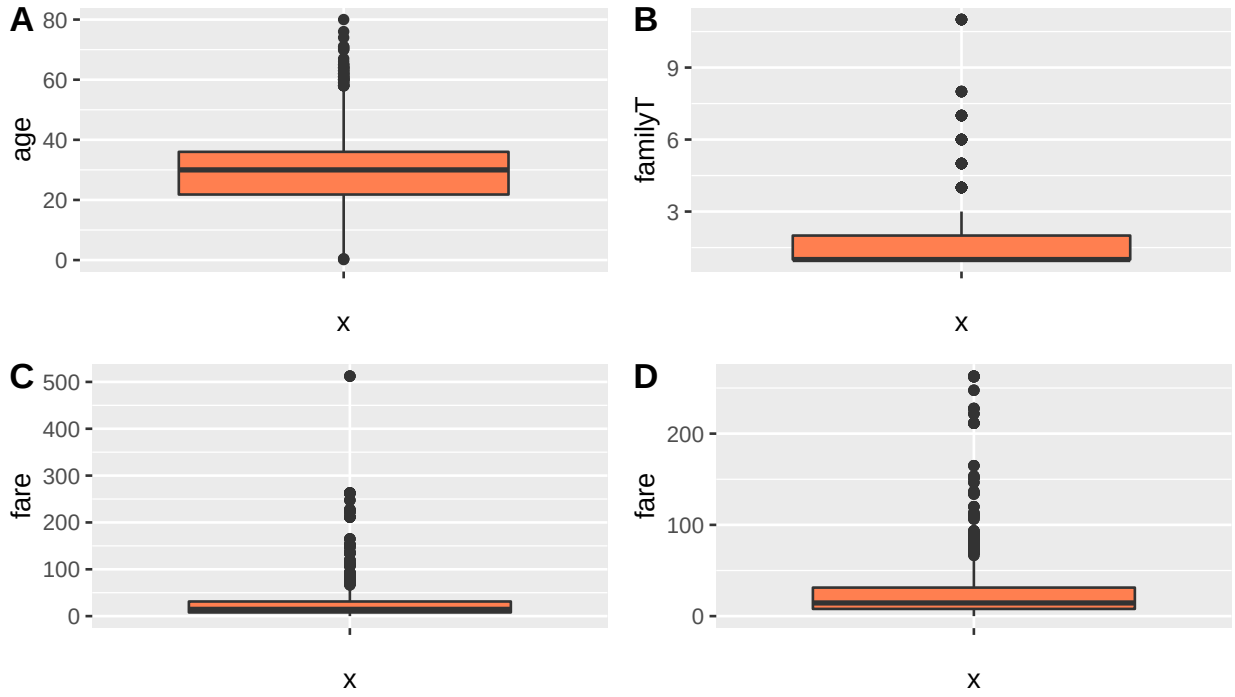
```
titanic$survived[titanic$survived == "1"] <- "Yes"
titanic$survived[titanic$survived == "0"] <- "No"
titanic$pclass <-as.factor(titanic$pclass)
```

Con estas modificaciones damos por concluido el pre-procesado. ### Analisis Exploratorio de Datos (EDA).

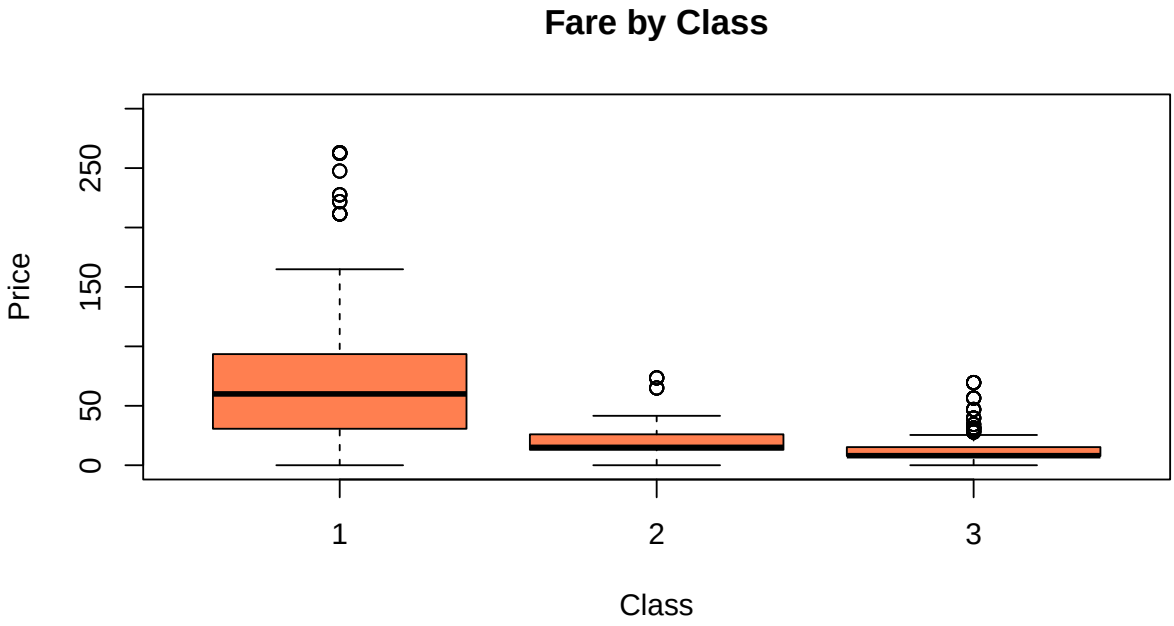
Nuestro dataset tiene ahora 12 variables: - Numericas: Age, fare,familyT. - Categoricas : pclass, sex, embarked, survived, title, family - Otras: passenger_id, name,ticket. Comenzaremos con las variables numericas, para ello obtendremos los boxplot de las mismas para ver su distribucion y revisar la existencia de outliers.

Variables numericas.

```
# Boxplot para variables numericas.
edad_boxplot <-ggplot(titanic) + aes(x= "",y=age)+geom_boxplot(fill ="coral")
familyT_boxplot <-ggplot(titanic) + aes(x= "",y=familyT)+geom_boxplot(fill ="coral")
fare_boxplot <-ggplot(titanic) + aes(x= "",y=fare)+geom_boxplot(fill ="coral")
# Tratamiento outliers para la variable fare
fare_outlier <- titanic$passenger_id[titanic$fare >=500]
# calculamos la media de los precios de los billetes de primera.
mean_first_class <- round(mean(titanic$fare[(titanic$pclass == "1")]),2)
titanic$fare[which(titanic$fare >500)] <- mean_first_class
# Obtenemos el boxplot con las modificaciones.
fare_boxplot2 <-ggplot(titanic) + aes(x= "",y=fare)+geom_boxplot(fill ="coral")
ggarrange(edad_boxplot,familyT_boxplot,fare_boxplot,fare_boxplot2, labels=c("A","B","C","D"),ncol=2,nrou
```




```
# Calculamos la distribucion de los precio de los billetes por clase
  boxplot(fare ~ pclass,data = titanic, col = "coral",
  ylab = "Price", xlab = "Class", ylim=c(0,300),
  main = "Fare by Class")
```



El recorrido intercuartilico de la edad se centra entre los 20 y los 40 años. El rango de edades de los pasajeros esta dentro de lo posible, igual que para el rango de miembros de una determinada familia, no realizamos actuacion alguna para ellas. Esta documentado que viajaba una familia de 11 miembros,

https://www.bbc.com/news/uk-england-cambridgeshire-17596264. De hecho son famosos para los analistas de este dataset por perecer la familia completa.

En el precio de los billetes observamos unos outlier muy significativo en la zona de 500.

Obtenemos los id de los pasajeros y vemos que son pasajeros de primera clase embarcados en Cherburgo.

Calculamos la media de los precios de billetes de primera y sustituimos el valor.

Tras eso adjuntamos el boxplot resultado etiquetado como D en el que claramente puede compararse con el original C y ver que las modificaciones realizados corrigen la situacion.

El precio del billete esta intimamente relacionado con la clase en la que se viajaba. Obtenemos los boxplot de fare y pclass.

Resulta sorprendente el hecho de que en las tres clases tenemos pasajeros con precio del billete cero.

La White Star, la empresa propietaria del barco tenia a parte de la tripulacion inscrita como viajeros y esta es la razon de dichos valores. La orquesta de barco estaba embarcada en segunda y tercera clase, asi como un reten de trabajadores del astillero donde se fabrico, para resolver cualquier eventualidad que pudiese surgir en este primer viaje.

Visualmente podemo apreciar que la mayoria de los billetes de tercera clase estan entre los 10 y 15 dolares, los de segunda entre 15 y 30 y los de primera entre 30 y 95, siendo esta clase la que mas dispersion presenta, ya que las comodidades contratadas podian variar mucho.

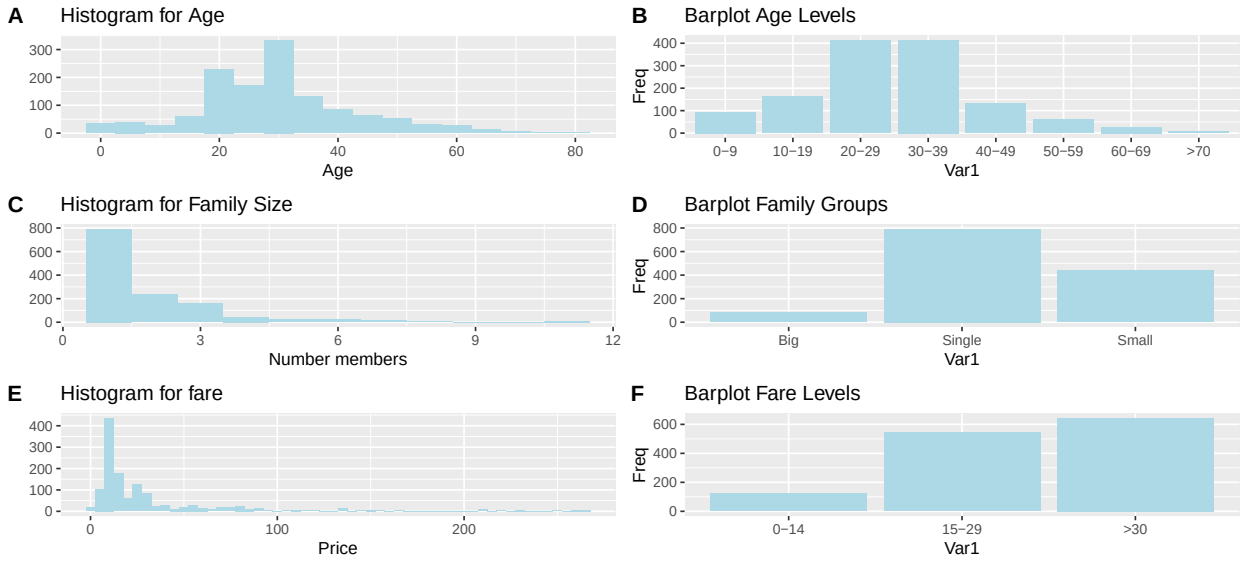
Es muy curioso la presencia de outliers en la tercera clase que se solapan en niveles de precio correspondientes a la primera clase.

Hecho esto vamos a obtener los histogramas donde veremos las distribuciones de las variables. La variable edad procederemos a discretizarla y la añadimos a los boxplot.

```
# Histogramas de las variables numericas.
age_histogram<-
qplot(titanic$age,geom="histogram",binwidth=5,
      main="Histogram for Age", xlab="Age", fill=I("lightblue"))
family_size_histogram<-
qplot(titanic$familyT,
      geom="histogram",binwidth=1,  main="Histogram for Family Size",
      xlab="Number members", fill=I("lightblue"))
fare_histogram<-
qplot(titanic$fare,
      geom="histogram",binwidth=5,
      main="Histogram for fare", xlab="Price", fill=I("lightblue"))
# Discretizacion de la variable Age.
titanic$age_levels <- cut(titanic$age, breaks = c(0,10,20,30,40,50,60,70,100),
                          labels =c("0-9","10-19","20-29","30-39","40-49",
                                    "50-59","60-69",">70"))
# Discratizacion variable Fare:
titanic$fare_levels <- cut(titanic$age, breaks = c(0,15,30,500),
                          labels =c("0-14","15-29",">30"))
# Barplot
age_levels_barplot <-ggplot(as.data.frame(table(titanic$age_levels))
                            ,aes(x=Var1, y=Freq,fill=I("lightblue"), xlab="Age")
                    )+geom_col() + ggtitle("Barplot Age Levels")

family_group_barplot <-ggplot(as.data.frame(table(titanic$family)),
                              aes(x=Var1, y=Freq,fill=I("lightblue"), xlab="Age")
                              )+geom_col() + ggtitle("Barplot Family Groups")
```

```
fare_levels_barplot <-ggplot(as.data.frame(table(titanic$fare_levels)),
                             aes(x=Var1, y=Freq,fill=I("lightblue"), xlab="Age")
                              )+geom_col() + ggtitle("Barplot Fare Levels")
ggarrange(age_histogram,age_levels_barplot,family_size_histogram,
          family_group_barplot, fare_histogram,fare_levels_barplot,
          labels=c("A","B","C","D","E","F"),ncol=2,nrow=3)
```



Comprobacion de la normalidad y homogeneidad de la varianza.

Observamos que las distribuciones tanto de fare como como de family_size estan sesgadas a la derecha, por lo que podemos descartar incluso visualmente que sean distribuciones normales. La distribuciond de age tampoco tiene la forma de la campana normal. No obstante comprobaremos la normalidad mediante la prueba de Anderson-Darling

```
# variable Age.
ad.test(titanic$age)
```

```
##
##  Anderson-Darling normality test
##
## data:  titanic$age
## A = 11.505, p-value < 2.2e-16
```

```
# variable Family.
ad.test(titanic$familyT)
```

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

```
# variable Family.
ad.test(titanic$fare)
```

```
##
##  Anderson-Darling normality test
##
## data:  titanic$fare
## A = 171.65, p-value < 2.2e-16
```

En todos los casos el p-valor es inferior a 0.05, por lo que podemos rechazar la hipotesis nula y afirmar que las tres distribuciones de datos que no sigan una distribucion normal.

Comprobamos ahora, la existencia de homogeneidad de la varianza para cada una de estas variables respecto a la variable survived, empleando para ello la prueba de Bartlett:

```
# Variable Age.
bartlett.test(titanic$age, titanic$survived)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  titanic$age and titanic$survived
## Bartlett's K-squared = 8.6479, df = 1, p-value = 0.003274
```

```
# Variable Family.
bartlett.test(titanic$familyT, titanic$survived)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  titanic$familyT and titanic$survived
## Bartlett's K-squared = 51.933, df = 1, p-value = 5.743e-13
```

```
# Variable Fare.
bartlett.test(titanic$fare, titanic$survived)
```

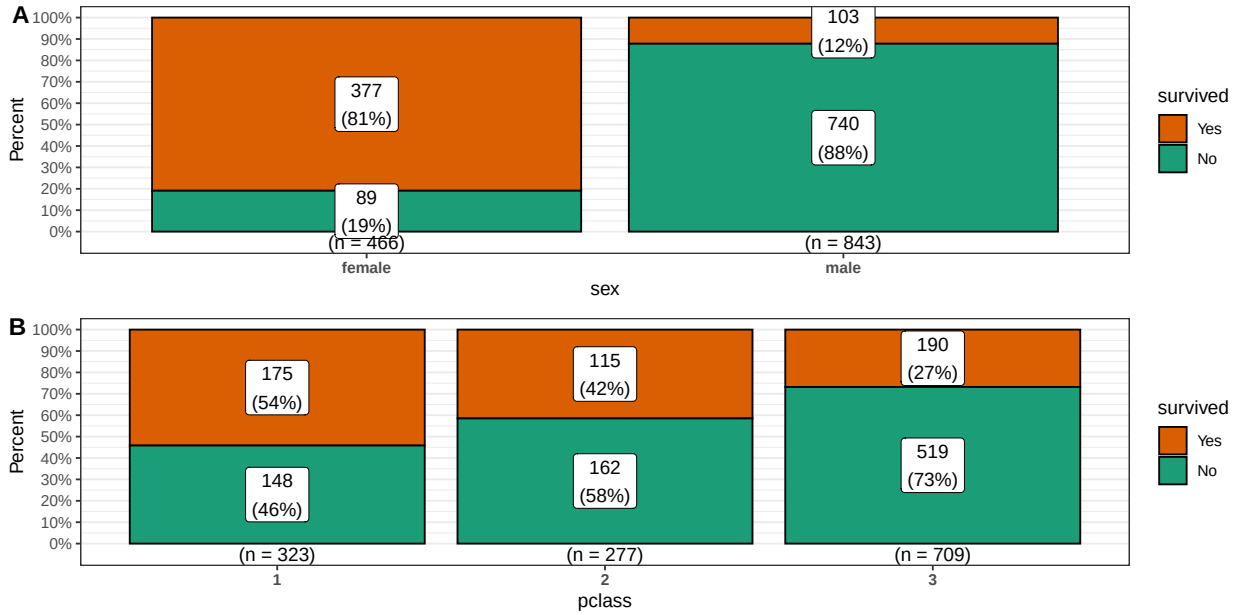
```
##
##  Bartlett test of homogeneity of variances
##
## data:  titanic$fare and titanic$survived
## Bartlett's K-squared = 138.64, df = 1, p-value < 2.2e-16
```

Todos los valores p-value son inferiores a 0.05, por lo que podemos rechazar la hipotesis nula, que establece que las varianzas son iguales, concluyendo que no existe homogeneidad en la varianza respecto a la variable survived.

Variables categoricas.

En primer lugar hay que mencionar que fallecieron 829 pasajeros un 63% y sobrevivieron 480 el 37%.

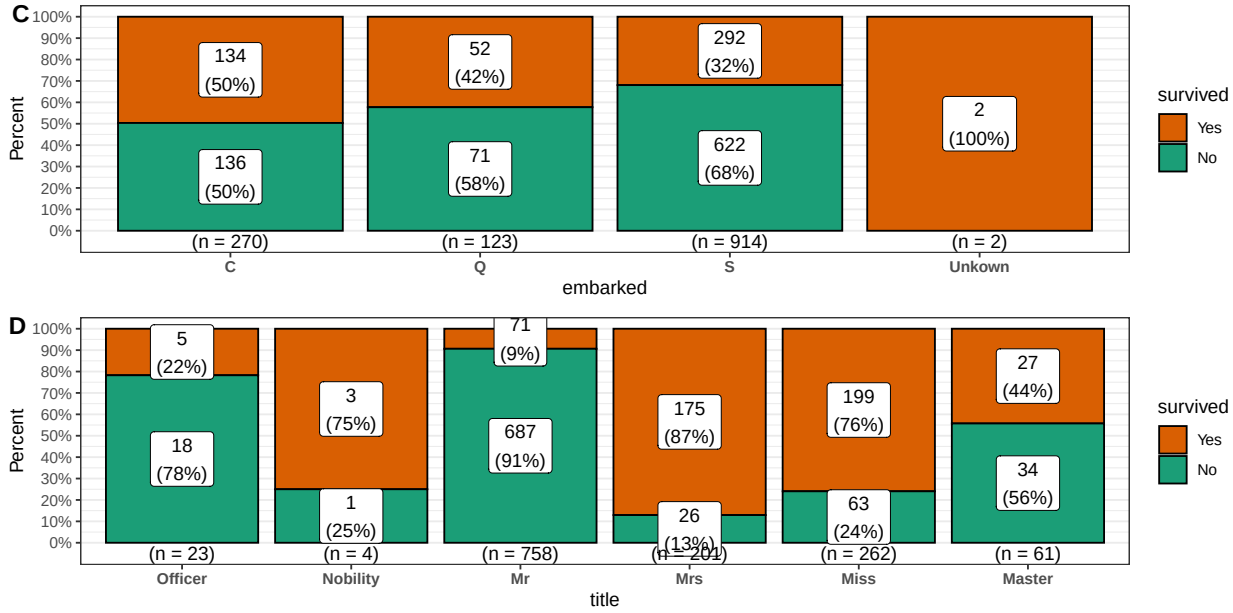
```
survived_by_sex <-CGPfunctions::PlotXTabs2(titanic,sex,survived,
                  data.label = "both",results.subtitle = FALSE )
survived_by_pclass <-CGPfunctions::PlotXTabs2(titanic,pclass,survived,
                    data.label = "both",results.subtitle = FALSE )
survived_by_embarked <-CGPfunctions::PlotXTabs2(titanic,embarked,survived,
                      data.label = "both",results.subtitle = FALSE )
survived_by_title <-CGPfunctions::PlotXTabs2(titanic,title,survived,
                      data.label = "both",results.subtitle = FALSE )
survived_by_familyT <-CGPfunctions::PlotXTabs2(titanic,familyT,survived,
                      data.label = "both",results.subtitle = FALSE )
survived_by_family <-CGPfunctions::PlotXTabs2(titanic,family,survived,
                      data.label = "both",results.subtitle = FALSE )
survived_by_age_levels <-CGPfunctions::PlotXTabs2(titanic,age_levels,survived,
                        data.label = "both",results.subtitle = FALSE )
survived_by_age <-CGPfunctions::PlotXTabs2(titanic,age,survived,
                  data.label = "both",results.subtitle = FALSE )
ggarrange(survived_by_sex, survived_by_pclass,labels=c("A ","B"),ncol=1,nrow=2)
```



En el grafico A observamos que la tasa de supervivencia entre las mujeres fue mucho mayor que entre los hombres. En el caso de las mujeres sobreviven un 81% frente a un 12% de los hombres.

En el grafico B tenemos la supervivencia en funcion de la clase en la que se viajaba. Facilmente podemos apreciar que es la tercera clase la que tienen peor tasa de supervivencia. Mientras que en la clases primera y segunda la probabilidad de perecer es del uno de cada dos viajeros, en el caso de tercera clase ascienda a dos de cada tres.

```
ggarrange(survived_by_embarked, survived_by_title,labels=c("C ","D"),ncol=1,nrow=2)
```



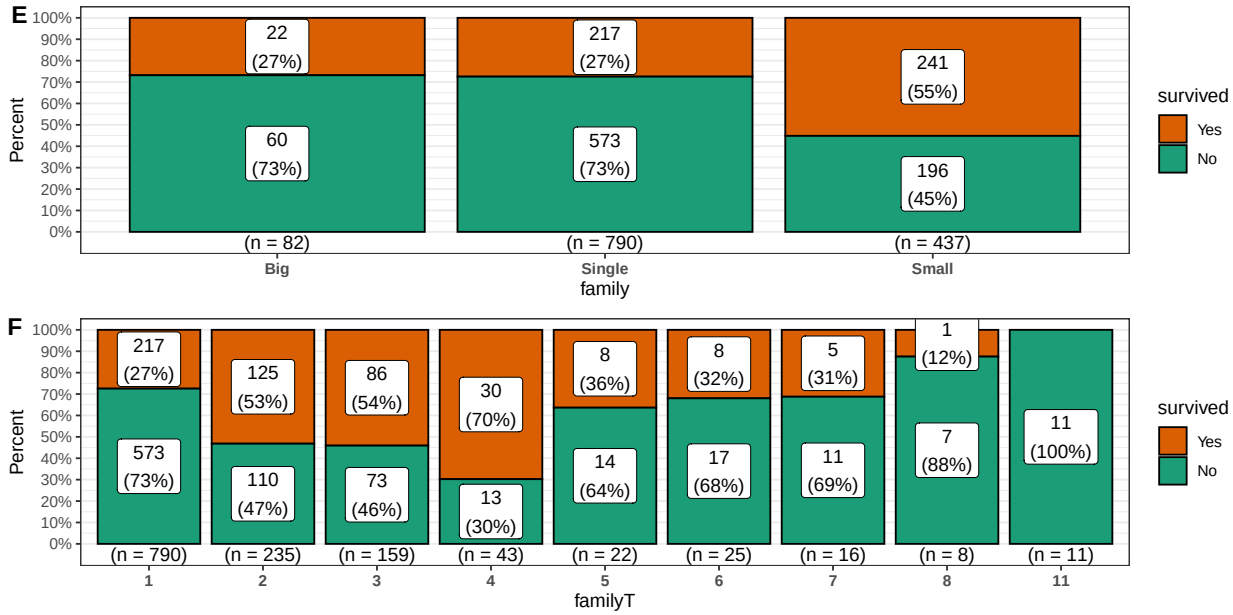
El barco parte de Southampton(S), curiosamente es ahi donde tenemos la menor posibilidad de supervivencia. Tras ello se dirige a Cherburgo (C) en Francia es desde este punto de embarque donde la probabilidad de supervivencia es la mayor, tras ello se dirigio a Queenstown (Q) en Irlanda siendo esta la menor de todas.

En cuanto a los titulos tenemos altas probabilidades de supervivencia entre las mujeres Mrs,Miss con una tasa de un 75 o superior. Fuera de ellos unicamente los Nobility tienen una tasa supervivencia similar.

En cambio los Mr hombres son mucha diferencia el colectivo con peor probabilidad de supervivencia, asi como los Officer que son los militares.

En cuanto a los niños podemos ver que, visto ambos extremos, tienen una probabilidad de supervivencia media.

```
ggarrange(survived_by_family , survived_by_familyT ,labels=c("E ","F"),ncol=1,nrow=2)
```



En cuanto a la situacion familiar en el grafico E tenemos el factor de familias que habiamos reducido a tres

niveles para facilitar la lectura.

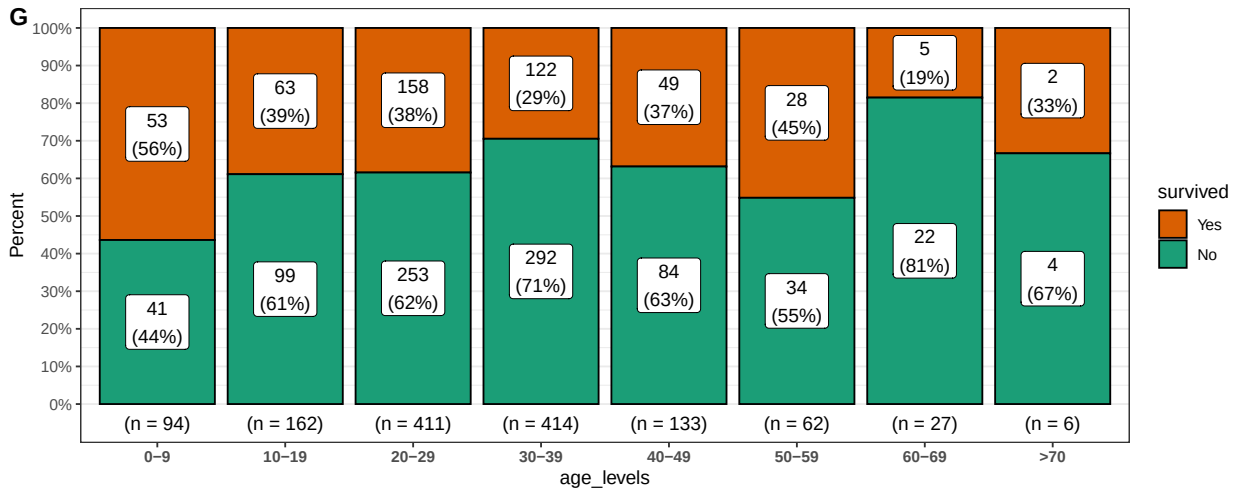
Curiosamente la gente soltera es la que tiene peores probabilidades de sobrevivir. Son las familias pequeñas de menos de 5 componentes los que con diferencia tienen mejor posibilidad de sobrevivir con un 55%

En el caso de las familias grandes, tienen similar tasa de supervivencia que los solteros.

Dentro de las familias pequeñas,las de tamaño maximo, bajo nuestra definicion, de 4 personas son las que presentan mejor probabilidad de supervivencia.

Para las familias grandes la distribucion es muy similar a lo largo de todo su rango. Hay que mencionar como aparece el caso que vimos al tratar el outlier de esta variable en los 11 miembros, que fallece toda la familia al completo.

```
ggarrange(survived_by_age_levels,labels=c("G","H"),ncol=1,nrow=1)
```




```
write.csv(titanic,"c:\\pruebas\\titanic_processed.csv") # obtenemos archivo final
```

Pruebas para comparar los grupos de datos.

Ahora que tenemos una idea mas aproximada de que factores contribuyeron en mayor medida a la supervivencia de un pasajero procedemos a ver si hay relacion de dependencia estadistica significativa entre las variables categoricas y la supervivencia o no del viajero.

Para ello empleamos tablas de contigencia para una prueba Chi-Cuadrado.

```
generarTablaContingencia <- function(principal,grupo,titulo){
tablaCont <- table(principal,grupo)
chi_cuadrado <- chisq.test(tablaCont)
titulo <- paste(titulo,". p-value: ", chi_cuadrado$p.value)
tablaCont <- ggplot() + annotation_custom(tableGrob(tablaCont)) + labs(title = titulo)}
cont1 <- generarTablaContingencia (titanic$survived,titanic$sex,'Sex vs Survival')
cont2 <- generarTablaContingencia (titanic$survived,titanic$pclass,'Pclass vs Survival')
cont3 <- generarTablaContingencia (titanic$survived,titanic$embarked,'Embarked vs Survival')
cont4 <- generarTablaContingencia (titanic$survived,titanic$title,'Title vs Survival')
cont5 <- generarTablaContingencia (titanic$survived,titanic$age_levels,'Age range vs Survival')
grid.arrange(cont1, cont2, cont3, cont4, cont5, nrow=5)
```

Sex vs Survival . p−value: 5.95438583970967e−134

	female	male
No	89	740
Yes	377	103

Pclass vs Survival . p−value: 4.69451185387565e−17

	1	2	3
No	148	162	519
Yes	175	115	190

Embarked vs Survival . p−value: 2.6219905008172e−07

	C	Q	S	Unkown
No	136	71	622	0
Yes	134	52	292	2

Title vs Survival . p−value: 8.06679199740111e−137

	Officer	Nobility	Mr	Mrs	Miss	Master
No	18	1	687	26	63	34
Yes	5	3	71	175	199	27

Age range vs Survival . p−value: 4.69435998093327e−05

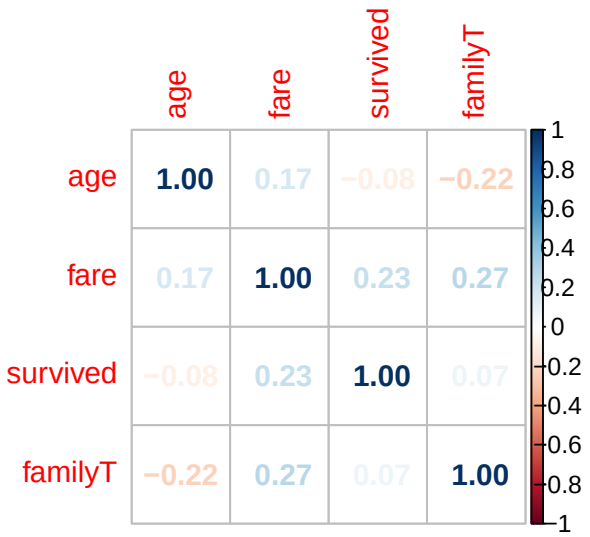
	0−9	10−19	20−29	30−39	40−49	50−59	60−69	>70
No	41	99	253	292	84	34	22	4
Yes	53	63	158	122	49	28	5	2

En cada tabla tenemos un p-valor, para todas las variables es inferior a 0.05, con lo cual podemos rechazar la hipotesis nula de que no existe dependencia entre cada una de las variables y Survived. Por tanto podemos concluir que si que existe una dependencia significativa entre las variables y survived.

Correlacion de variables.

Examinamos la dependencia entre variables numericas con una matriz de correlacion en la que incluiremos la variable survived previo paso a variable numerica.

```
titanic$survived[titanic$survived == "Yes"] <- "1"
titanic$survived[titanic$survived == "No"]  <- "0"
titanic$survived <- as.numeric(titanic$survived)
# Seleccionamos las variables numericas que tenemos:
titanic_num <- titanic %>% select(where(is.numeric))
# Creamos la matriz de correlacion.
corr_data <- cor(titanic_num)
corrplot(corr_data, method="number")
```



Todas las variables tienen una relacion muy debil. En el caso de age y fare es positiva y unicamente esta ultima es un poco mas elevada. Para el tamaño de la familia y edad es negativa.

Regresion lineal.

Modelo 1 - Modelo Sex

En primer lugar hemos visto que dentro del pasaje,la cifra de supervivencia fue mayor entre las mujeres, razon por la cual empezaremos con esta variable un primer modelo.

```
# Generamos modelo.
gender_model <- lm(survived ~ sex, data = titanic)
gender_model
```

```
##
## Call:
## lm(formula = survived ~ sex, data = titanic)
##
## Coefficients:
## (Intercept)      sexmale
##      0.8090      -0.6868
```

```
# Predecimos la supervivencia.
titanic$survived_gender_model <- round(predict(gender_model,titanic))
# Calculamos la proporcion de observacion correctamente clasificadas.
nrow(titanic[titanic$survived_gender_model == titanic$survived,]) / nrow(titanic)
```

```
## [1] 0.8533231
```

Obtenemos una precision del 85%, con una unica variable es un buen valor.

Modelo 2 - Modelo Sex-Pclass

En el EDA apreciamos mayor supervivencia en las clases 1 y 2, razon por la que añadimos esta variable.

```r
# Generamos modelo.
gender_class_model <- lm(survived ~ sex + pclass , data = titanic)
gender_class_model
```

```
##
## Call:
## lm(formula = survived ~ sex + pclass, data = titanic)
##
## Coefficients:
## (Intercept)      sexmale      pclass2      pclass3
##     0.91150     -0.66712     -0.08451     -0.17964
```

```r
# Predecimos la supervivencia.
titanic$survived_gender_class_model <- round(predict(gender_class_model,titanic))
# Calculamos la proporcion de observacion correctamente clasificadas.
nrow(titanic[titanic$survived_gender_class_model == titanic$survived,]) / nrow(titanic)
```

```
## [1] 0.8533231
```

Vemos que añadir esta variable no contribuye a mejorar la precision del modelo que se mantiene en un 85%.

Modelo 3 - Añadimos title.

```r
# Generamos modelo.
complete_model_1 <- lm(survived ~ sex + pclass + title,  data = titanic)
complete_model_1
```

```
##
## Call:
## lm(formula = survived ~ sex + pclass + title, data = titanic)
##
## Coefficients:
##   (Intercept)        sexmale        pclass2        pclass3  titleNobility
##      0.864081      -0.635919      -0.088365      -0.184466       0.203878
##       titleMr       titleMrs      titleMiss    titleMaster
##     -0.007983       0.091756       0.017935       0.366477
```

```r
# Predecimos la supervivencia.
titanic$survived_complete_model_1 <- round(predict(complete_model_1 ,titanic))
# Calculamos la proporcion de observacion correctamente clasificadas.
nrow(titanic[titanic$survived_complete_model_1  == titanic$survived,]) / nrow(titanic)
```

```
## [1] 0.8624905
```

Modelo 4 - Añadimos age y fare

```r
# Generamos modelo.
complete_model_2 <- lm(survived ~ sex + pclass + title + age +  fare  ,  data = titanic)
complete_model_2
```

```
##
## Call:
## lm(formula = survived ~ sex + pclass + title + age + fare, data = titanic)
##
## Coefficients:
##   (Intercept)        sexmale        pclass2        pclass3  titleNobility
##     0.9908734     -0.6299684     -0.1242022     -0.2268034      0.1817625
##       titleMr       titleMrs      titleMiss    titleMaster            age
##    -0.0291537      0.0857063     -0.0169853      0.2939652     -0.0022554
##          fare
##    -0.0002825
```

```
# Predecimos la supervivencia.
titanic$survived_complete_model_2 <- round(predict(complete_model_2 ,titanic))
# Calculamos la proporcion de observacion correctamente clasificadas.
nrow(titanic[titanic$survived_complete_model_2  == titanic$survived,]) / nrow(titanic)
```

```
## [1] 0.8624905
```

La variable que mas poder explicativo tiene es el sexo esa variable por si sola nos da una precision del 85%. Si lo que queremos es minimizar el numero de variablesy por tanto primamos la sencillez, este seria el que escogeriamos. Añadiendo pclass y title alcanzariamos el valor de 86% y este seria el modelo que mayor potencia tendria, esto es, elegiriamos este si, lo que buscamos es maximizar la precision empleando el menor numero de variables posible. Añadir mas variable no merece la pena ya que aumenta el poder explicativo del modelo.

Conclusiones finales.

El presente caso practico se desarrolla sobre el dataset que contiene la lista de pasajeros del buque HMS Titanic, hundido en su viaje inaugural en 1912. El objetivo de esta practica es determinar que factores contribuian a incrementar las posibilidades de supervivencia. Se nos proporcionan 3 archivos que contienenla informacion de dicho pasaje, se procede a integrarlos. Esto es, por una parte se integra la informacion necesaria de la variable survived en el archivo test con la informacion del archivo baseline. Efectuado esto se unen los archivos train y test, ya tenemos un unico archivo de trabajo. Comprobamos que no hay datos duplicados , asi como valores NAs y cadenas vacias. Dependiendo de los casos se han eliminado, previa justificacion, variables o se han realizado imputaciones en dichas variables. Se procede a la limpieza de los datos, se discretizan variables, se crean nuevas derivadas que facilitan el estudio unificando la informacion que contenian por separado.

En el analisis exploratorio observamos:

El principal factor que explica la supervivencia del pasaje es el sexo el 81% de las mujeres sobrevive frente al 12% de hombres, algo que se haya refrendado por el atributo derivado title en el que podemos encontrar esta misma informacion desagregada por estado civil. Tras esto la clase en la que se viajaba era el siguiente factor clave en la supervivencia. En tanto que un pasajero de primera y segunda tenia una probabilidad de supervivencia de 54% y 42% respectivamente, esta bajaba al 27% en caso de viajar en tercera. El pasaje del Titanic era mayoritariamente gente joven de entre 20-40 años y soltera, siendo estos grupos los que mayor mortalidad registran. En el caso de familias pudimos ver tanto en la variable numerica como en el factor que creamos que las posibilidades de supervivencia en los casos de familias pequeñas, eran mauores que en los casos de familias grandes. En cuanto a supervivencia por edades destacar, la elevada tasa, dentro de los resultados globales, de supervivencia de los niños de hasta 10 años. Las principales variables cuantitativas que identificamos en el paso anterior, age, fare, family y podemos afirmar que ninguna seguia una distribucion normal.

Comprobamos la existencia de homogeneidad de la varianza de dichas variables respecto a survived mediante la prieba de Bartlett y podemos afirmar que no existe homogeneidad en la varianza respecto a la variable survived. Se calculan tablas de contingencia para prueba Chi-Cuadrado y de este modo ver si hay dependencia estadistica significativa entre las variables categoricas y survived. El resultado de la misma nos indica que si que existe una dependencia significativa entre esta variables y survived. Hecho esto calculamos la matriz de correlacion para ver dichas dependencias. El resultado que obtenemos es bastante esclarecedor. Existiendo relacion, esta es muy debil en todos los casos. Es positiva para age y fare y negativa para tamaño de la familia y edad, insistiendo en el hecho de comun de que todas son muy debiles. Por ultimo procedemos a calcular modelos de regresion lineal a partir de toda la informacion recopilada anteriormente. Comenzamos con un modelo muy sencillo empleando unicamente la variable sex, que nuestro EDA mostraba como mas determinante a la hora de mejorar la probabilidad de supervivencia. Vemos que con este modelo la precision que obtenemos clasificando observaciones es de un 85 %, siendo el mejor de los modelos que hemos obtenido si lo que primamos es la sencillez del modelo unido a un ratio razonable. Añadimos mas variables y obtenemos el modelo de mayor poder explicativo añadiendo a sex, pclass y title conun 86% de precision clasificando observaciones. Este seria el modelo que elegiriamos si lo que buscamos es optimizar el poder predictor del modelo con el minimo de variables posible. Por ultimo añadimos el resto de variables cuantitativas que tenemos, no experimentado mejora alguna en el poder predictor del modelo. Por tanto, el mejor modelo posible de los que hemos probado seria el de tres variables sex, pclass y title.

Bibliografia.

- Materiales de la asignatura
- Encyclopedia Titanica
  https://www.encyclopedia-titanica.org/
- "Titanic: a deeper look on family size"
  https://www.kaggle.com/lperez/titanic-a-deeper-look-on-family-size
- "Titanic simple rf with name and age features"
  https://www.kaggle.com/ianwells/titanic-simple-rf-with-name-and-age-features
- "Titanic Tragedy: Exploratory Data Anylisis". https://mohitatgithub.github.io/2018-03-08-Titanic-Tragedy-Exploratory-Data-Analysis/
- "Looking for Survivors with Titanic Data Analysis" https://jasonicarter.github.io/survival-analysis-titanic-data/