Tipologia y ciclo de vida de los datos: Practica 2 - Limpieza y analisis de datos

Autor: Joan Miquel Forteza Fuster

24/05/2020

---

1 Titanic: Analisis de datos y limpieza

Dato de contraste: Los diferentes archivos y resoluciones de la práctica se pueden encontrar en el siguiente enlace.
https://github.com/viciony/Tipologia-PRAC2

1.1 Presentación

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

1.2 Competencias

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

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

1.3 Objetivos

Los objetivos concretos de esta práctica son:

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

1.4 Descripcion de la PAC a realizar

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. Porque es importante y qué pregunta / problema pretende responder?
2. Integración y selección de los datos de interés a analizar.
3. Limpieza de los datos. 3.1. Los datos contienen ceros o elementos vacíos? Cómo gestionar estos casos? 3.2. Identificación y tratamiento de valores extremos.
4. Análisis de los datos. 4.1. Selección de los grupos de datos que se quieren analizar / comparar (planificación de los análisis a aplicar). 4.2. Comprobación de la normalidad y homogeneidad de la varianza. 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y del objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.
5. Representación de los resultados a partir de tablas y gráficas.
6. Resolución del problema. A partir de los resultados obtenidos, cuáles son las conclusiones? Los resultados permiten responder al problema?
7. Código: Hay que adjuntar el código, preferiblemente en I, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo prefiere, también puede trabajar en Python.

---

2 Desarrollo de la práctica

2.1 Descripcion del dataset

En este práctica trabajaremos con el Famoso Titánic conocido por casi todo el mundo, trabajaremos con un dataset que contiene los datos historicos del hundimiento del barco. Es un conjunto de datos muy usado por todo el mundo, ya que es un dataset sencillo y bastante facil de manipular para relizar las primeras prácticas con manipulación de datos. Ademas los datos que contiene el dataset son muy entendibles. El conjunto de datos proporciona un resemun del destino de los pasajeros a bordo del barco, clasificandolos segun clase economica, sexo, supervivencia y la edad del pasajero. Lo que se suele hacer con este conjunto de datos y muy típico, es crear modelos predictorios para saber si los pasajeros del conjunto de datos sobrevivieron o no al accidente.

Los ficheros csv que contienen el conjunto de datos son proporcionado por la página Web Kaggle, en el siguiente enlace. Esta página nos proporciona ficheros de entrenamiento y test, donde en dicho fichero de entrenamiento incluye la columna de classificacion de si sobrevive o no, en cambio en el de test no aparece dicha columna classificatorio, y esto es debido a que Kaggle lo proporciona asi, ya que realiza una competicion con los modelos que genera la gente a partir de estos datos. Se genera el modelo con el dataset de entrenamiento y se evalua con el dataset de test, de esta manera se obtiene una putuación de la certeza del modelo.

2.2 Integración y selección de los datos de interés a analizar

Lo primero de todo que haremos sera cargar el fichero CSV (Comma separated values) de entrenamiento que utilizaremos para nuestro análisis, estos ficheros contienen los datos separados por comas o a veces por punto y coma. Indicaremos también al cargar los datos que las cadenas de texto no las convierta automaticamente a factores, ya que se decidiran a lo largo de la práctica a través de el análisis de estos datos.

```
# Caragamos los paquetes que utilizaremos para realizar la práctica, y en concreto el paquete para u
sar las gráficas
library(dplyr)
library(ggplot2)

# Guardamos nuestro conjunto de datos dentro el dataset dtTitanic
dtTitanic <- read.csv('../DATASETS/train-titanic.csv', stringsAsFactors = FALSE)

# Observamos la estructura que tiene nuestro dataset
str(dtTitanic)
```

```
## '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" ...
```

Como podemos ver nuestro conjunto de datos contiene 891 filas en este fichero de entrenamiento. Podemos observar que disponemos de 12 variables en nuestro conjunto, 7 continuas y 5 clasificatorias, las cuales son:

- `PassengerId` : Identificador unico para cada pasajero
- `Survived` : Nos indica si sobrevivio o no (0 = No , 1 = Si)
- `Pclass` : Clase económica del pasajero (1 = primera, 2 = segunda, 3 = tercera)
- `Name` : Nombre del pasajero
- `Sex` : Sexo del pasajero
- `Age` : Edad de los pasajeros
- `SibSp` : Numero de hermanos en la embarcación
- `Parch` : Numero de padre/hijo en la embarcación
- `Ticket` : Numero de billete
- `Fare` : Precio del billete
- `Cabin` : Identificador de la cabina de la embarcación
- `Embarked` : Puerto donde el pasajero subio a la embarcación

A partir de estas variables realizaremos un análisis de la supervivencia de nuestros pasajeros en el conjunto de datos.

2.3 Limpieza de datos

2.3.1 Los datos contienen ceros o elementos vacíos? Cómo gestionar estos casos?

Primero de todo, analizaremos nuestros outliers, valores atípicos, nulls, etc.. que puede contener nuestro conjunto de datos. Lo siguiente que analizaremos son los valores vacios que tenemos.

```
# Datos estadísticos con valores atípicos
colSums(is.na(dtTitanic))
```

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

```
colSums(dtTitanic=="")
```

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

```
# A los valores vacios que tenemos en la variable Embarked les ponemos C
dtTitanic$Embarked[dtTitanic$Embarked==""]="C"

# Para los valores null de la variable de edad, realizaremos una mediana de nuestro conjunto de dato
s
dtTitanic$Age[is.na(dtTitanic$Age)] <- mean(dtTitanic$Age,na.rm=T)
```

El tema de factorizar las variables, requiere un previo análisis para ver que variables contienen pocas clases y realmente merece la pena factorizar, esto lo haremos mediante la función apply de R, que nos mostrará lo que buscamos. Despues de esto podremos factorizar las diferentes clases que veamos.

```
# Vemos la longitud de las variables
apply(dtTitanic,2, function(x) length(unique(x)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##         891           2           3         891           2          89
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           7           7         681         248         148           3
```

```
# Factorizamos las variables con pocas categorias
cols<-c("Survived","Pclass","Sex","Embarked")
for (i in cols){
  dtTitanic[,i] <- as.factor(dtTitanic[,i])
}

# Visualizamos la estructura del dataset
str(dtTitanic)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs
Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : 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   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
```

Como podemos ver en las variables tenemos conjuntos que tienen 7 o 8 distintas categorias pero no quiere decir que vayamos a factorizarlos, por que en el caso de una seria el numero de hermanos que tiene un pasajero y eso se puede repetir mucho su valor, pero en este caso solo se factorizaran variables con menor valor.

Las variables que se han factorizado han sido `Survived` con dos categorias, `Pclass` con 3 categorias, `Sex` con dos categorias y `Embarked` con 3 categorias.

También observaremos los valores que contienen el 0 en nuestro conjunto de datos

```
# Valores 0
colSums(dtTitanic==0)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         549           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##         608         678           0          15           0           0
```

Hay 4 variables que contienen valores a 0, pero estos valores no se consideraran como valores que sean malos para la análitica, ya que en el algunos casos como el de los hermanos y hijos y padres se puede valorar que no tiene, y en el caso de la variable Fare, podemos

considerar que esas personas viajaron de gratis a bordo de la embarcación, que con la siguiente instrucción podemos ver que fueron todo hombres que embarcaron desde el mismo puerto y iban solos, por lo tanto podia ser tripulación de refuerzo de la embarcación que fueron considerados como pasajeros o personas invitadas a la embarcación.

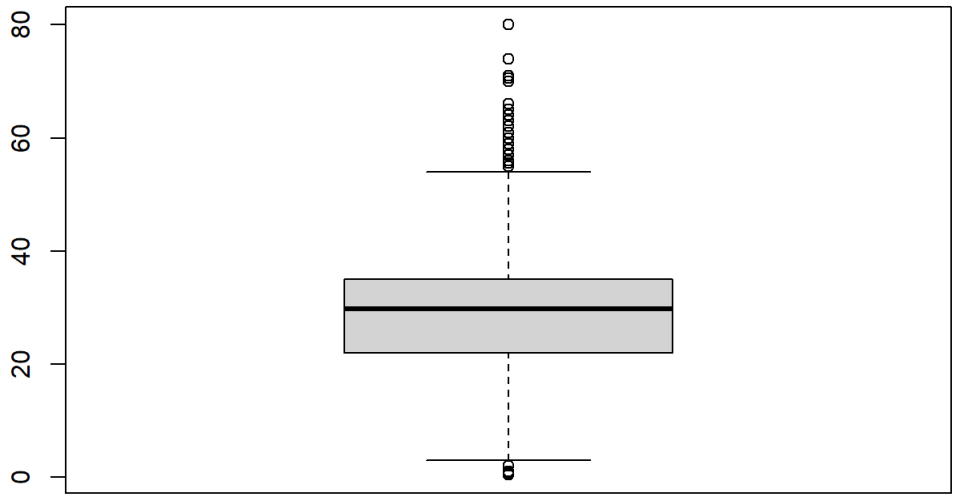
```
# Conjunto de valores a 0 de Fare
dtTitanic[dtTitanic$Fare==0,]
```

```
##     PassengerId Survived Pclass                               Name  Sex      Age
## 180         180        0      3              Leonard, Mr. Lionel male 36.00000
## 264         264        0      1            Harrison, Mr. William male 40.00000
## 272         272        1      3     Tornquist, Mr. William Henry male 25.00000
## 278         278        0      2      Parkes, Mr. Francis "Frank" male 29.69912
## 303         303        0      3  Johnson, Mr. William Cahoone Jr male 19.00000
## 414         414        0      2   Cunningham, Mr. Alfred Fleming male 29.69912
## 467         467        0      2            Campbell, Mr. William male 29.69912
## 482         482        0      2 Frost, Mr. Anthony Wood "Archie" male 29.69912
## 598         598        0      3              Johnson, Mr. Alfred male 49.00000
## 634         634        0      1    Parr, Mr. William Henry Marsh male 29.69912
## 675         675        0      2       Watson, Mr. Ennis Hastings male 29.69912
## 733         733        0      2             Knight, Mr. Robert J male 29.69912
## 807         807        0      1           Andrews, Mr. Thomas Jr male 39.00000
## 816         816        0      1                Fry, Mr. Richard male 29.69912
## 823         823        0      1  Reuchlin, Jonkheer. John George male 38.00000
##     SibSp Parch Ticket Fare Cabin Embarked
## 180     0     0   LINE    0              S
## 264     0     0 112059    0   B94        S
## 272     0     0   LINE    0              S
## 278     0     0 239853    0              S
## 303     0     0   LINE    0              S
## 414     0     0 239853    0              S
## 467     0     0 239853    0              S
## 482     0     0 239854    0              S
## 598     0     0   LINE    0              S
## 634     0     0 112052    0              S
## 675     0     0 239856    0              S
## 733     0     0 239855    0              S
## 807     0     0 112050    0   A36        S
## 816     0     0 112058    0  B102        S
## 823     0     0  19972    0              S
```

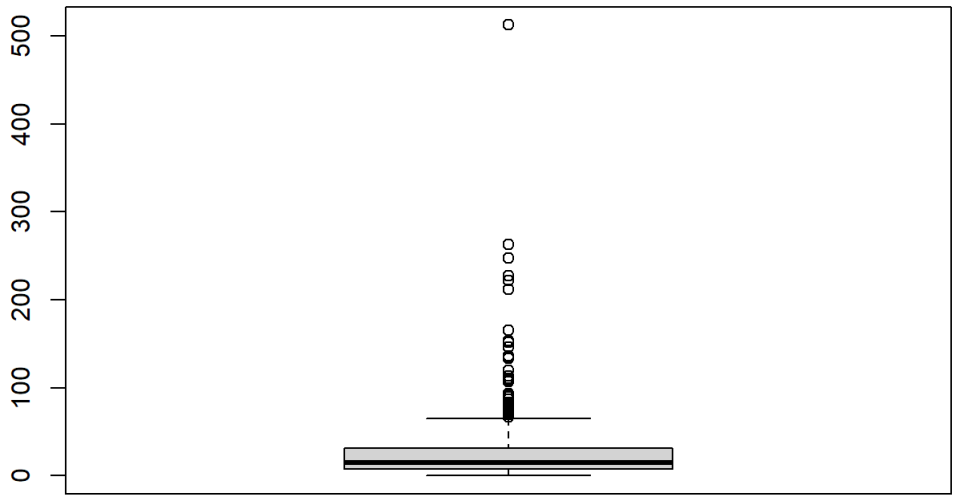
2.3.2 Identificación y tratamiento de valores extremos

En este apartado identificaremos los valores atípicos o outliers. Estos valores suelen representarse en diagramas de caja y son valores que estan en los extremos de estas mismas, o son muy grandes o muy pequeños para el conjunto de datos. Usaremos la función boxplot() de R para representar esos diagramas de caja y ver nuestros valores atípicos. Se utilizaran las variables no factorizadas y que son númericas para poder ver estos valores

```
boxplot(dtTitanic$Age)
```

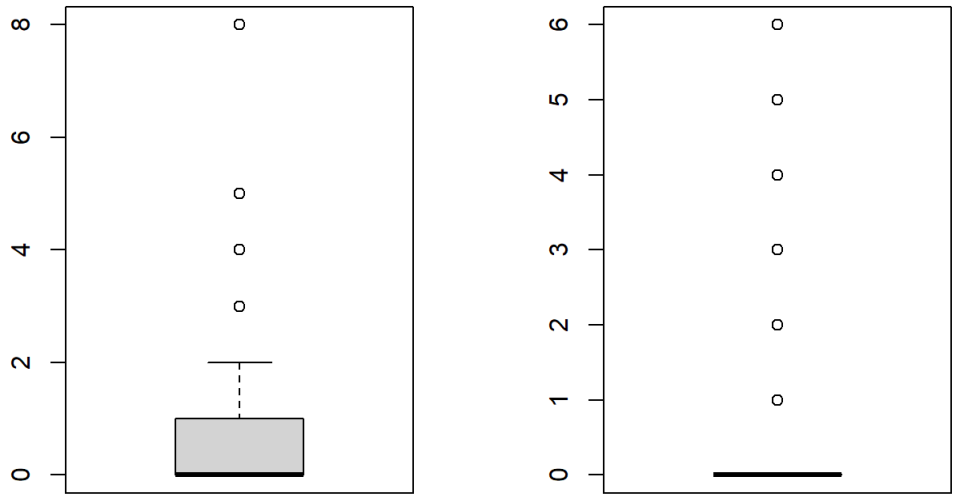



```
boxplot(dtTitanic$Fare)
```



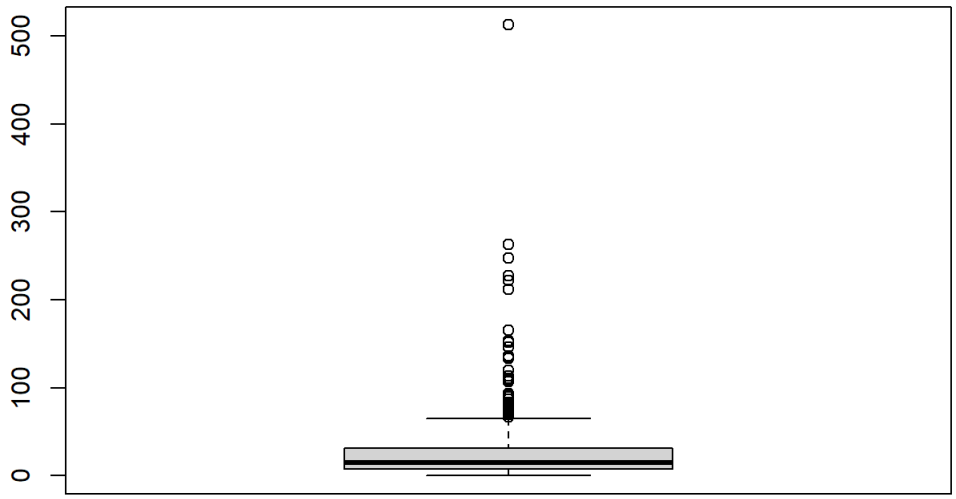
En este diagrama de caja vemos la edad de los pasajeros, los valores atípicos, són bebes o abuelos que estan entre los pasajeros por lo tanto los valores no se trataran y se consideraran buenos.

```
par(mfrow=c(1,2))
boxplot(dtTitanic$SibSp)
boxplot(dtTitanic$Parch)
```



En estos diagramas que juntamos son los diagramas que contienen el tamaño familiar, por lo tanto los valores atípicos se pueden considerar también por que es muy normal que haya familias a bordo numerosas o no.

```
boxplot(dtTitanic$Fare)
```



En este ultimo diagrama de caja podemos visualizar un valor atípico en el coste de los billetes que sera analizado mas adelante para observar detenidamente.

2.4 Analisis de los datos

En este apartado analizaremos los datos que sean de interes en nuestro conjunto de datos. Para eso crearemos una nueva variable también a partir de dos variables que podemos juntar, que son los nombres de padres y hermanos a bordo, de esta manera se podra hacer una analisis por numero de familiares a bordo.

```
# nueva variable nFamiliares
dtTitanic$nFamiliares <- dtTitanic$SibSp + dtTitanic$Parch + 1;

# Con la nueva variable queda así nuestro conjunto de datos
str(dtTitanic)
```

```
## 'data.frame':    891 obs. of  13 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs
Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : 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   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ nFamiliares: num  2 2 1 2 1 1 1 5 3 2 ...
```

En el apartado de limpeza de datos hemos detectado un valor atípico en la variable `Fare` mediante el diagrama de caja ,donde podemos ver un valor muy grande que se ha pagado por el billete de de embarcación. La cantidad del valor atípico es de 512.3292 y repetida en tres pasajeros distintos, por lo tanto miraremos que relación pueden tener estos tres pasajeros.

```
outliers_fare <- boxplot.stats(dtTitanic$Fare)$out
outliers_fare[outliers_fare>500]
```

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

```
dtTitanic[dtTitanic$Fare>500,]
```

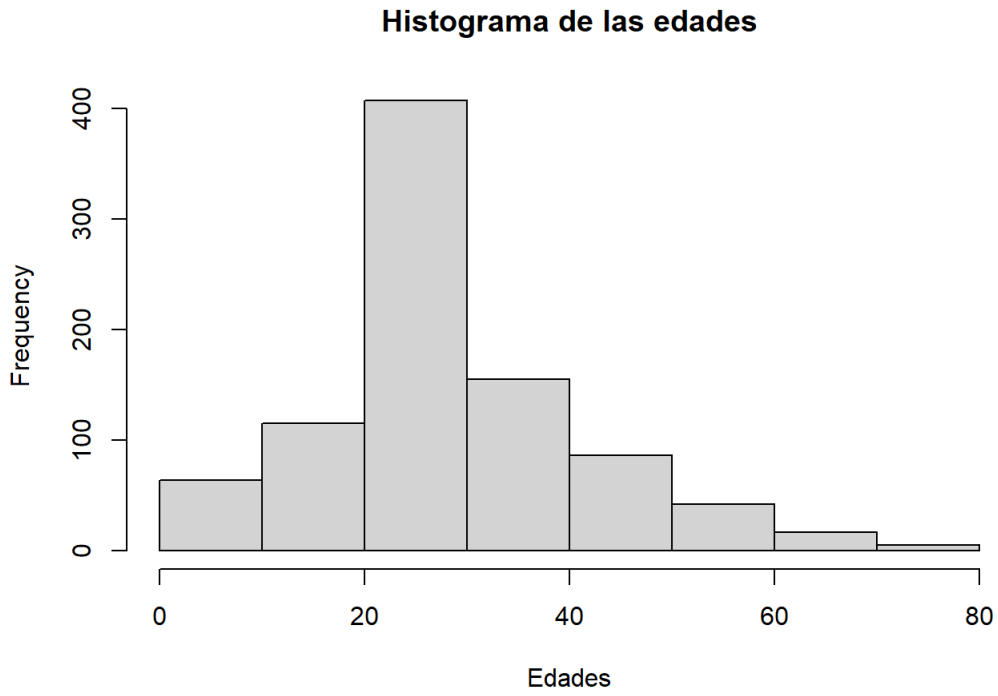
```
##     PassengerId Survived Pclass                                Name    Sex Age
## 259         259        1      1                    Ward, Miss. Anna female  35
## 680         680        1      1 Cardeza, Mr. Thomas Drake Martinez   male  36
## 738         738        1      1              Lesurer, Mr. Gustave J   male  35
##     SibSp Parch   Ticket     Fare       Cabin Embarked nFamiliares
## 259     0     0 PC 17755 512.3292                   C           1
## 680     0     1 PC 17755 512.3292 B51 B53 B55       C           2
## 738     0     0 PC 17755 512.3292        B101       C           1
```

En los anteriores datos podemos ver que los pasajeros no parecen tener una relación familiar entre ellos, sin embargo podemos observar que embaracaron en el mismo puerto, viajan en primera clase y tienen edades similares. Podemos ver que uno de ellos si tiene un familiar a bordo y los otros no, pero observando los datos podemos dar una conclusión de que el valor del billete puede ser por una clase especial de cabina o simplemente por un estatus jerarquico mas alto.

A continuación analizaremos las relaciones entre las diferentes variables de interes en nuestro conjuntos de datos y la variable que se marca como objetivo a predecir `Survived`.

Como primero analizaremos la relación entre la edad y la supervivencia, en el siguiente histrograma distrubiremos el conjunto de nuestra población según su edad. Lo que queremos comprobar es si a la hora de la supervivencia se tuvo en cuenta más los niños o los adultos, considerando niños todos los menores de 15 y los adultos a partir de esa edad para arriba. La media de la población de nuestro conjunto de datos es de 30 años.

```
hist(dtTitanic$Age, xlab = "Edades", main = "Histograma de las edades")
```



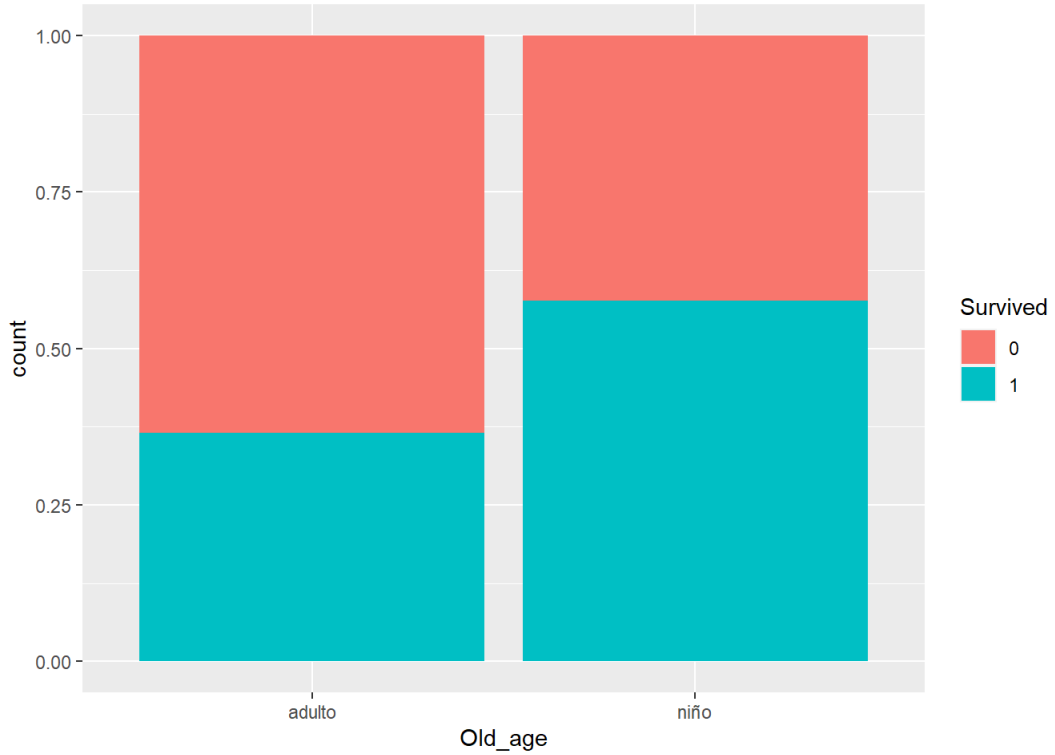
Generaremos una variable adicional factorial para clasificar y tener categorizados los adultos y los niños. Con esta variable podremos mostrar en los plots de una manera mas sencilla de analizar y respresentar que no por edades.

```
dtTitanic$Old_age <- 'adulto'
dtTitanic$Old_age[dtTitanic$Age < 15] <-'niño'

# Miramos cuantos niños y adultos tenemos en nuestro conjunto de datos
table(dtTitanic$Old_age)
```

```
##
## adulto   niño
##    813     78
```

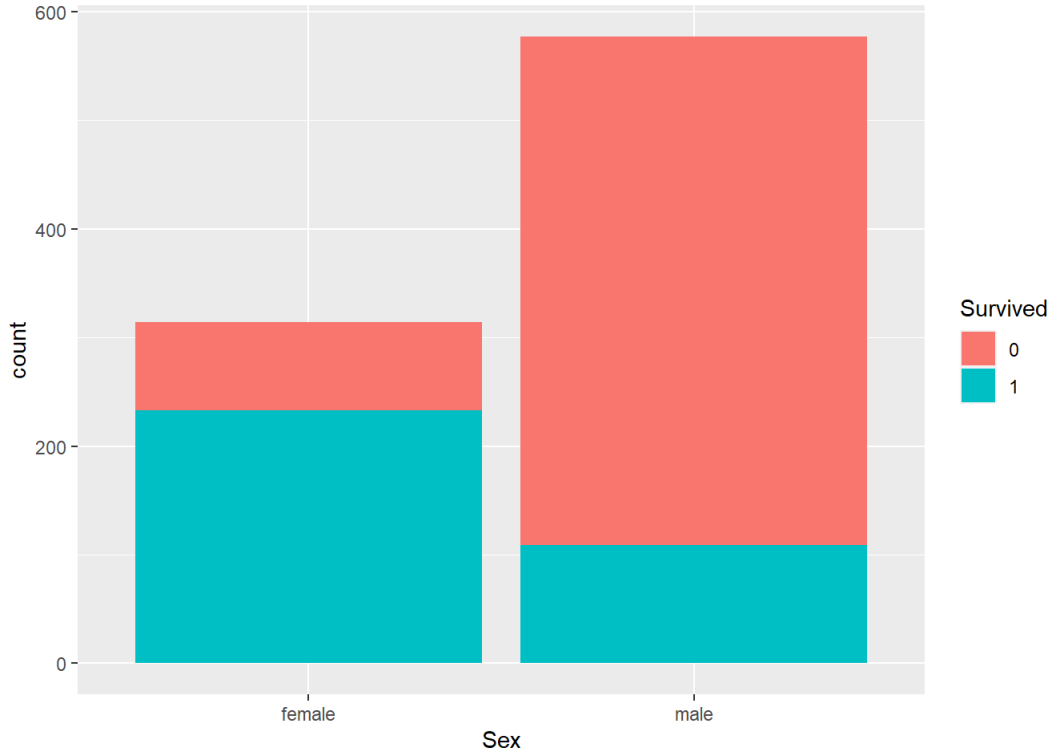
```
# Mostramos la relacion entre la variable nueva creada y la que se pretende como objetivo
ggplot(data = dtTitanic,aes(x=Old_age,fill=Survived))+geom_bar(position="fill")
```



Como podemos ver el porcentaje de niños es mas alto que el de adultos en supervivencia pero tambien el nombre de adultos es mas grande que el de niños.

En el siguiente plot coprobaremos la relación con otra variable destacable que es el sexo de los pasajeros y la variable objetivo survived.

```
ggplot(data=dtTitanic,aes(x=Sex,fill=Survived))+geom_bar()
```




```
tabla_comparativa<-table(dtTitanic$Sex,dtTitanic$Survived)
for (i in 1:dim(tabla_comparativa)[1]){
    tabla_comparativa[i,]<-tabla_comparativa[i,]/sum(tabla_comparativa[i,])*100
}

tabla_comparativa
```

```
##
##                 0        1
##   female 25.79618 74.20382
##   male   81.10919 18.89081
```

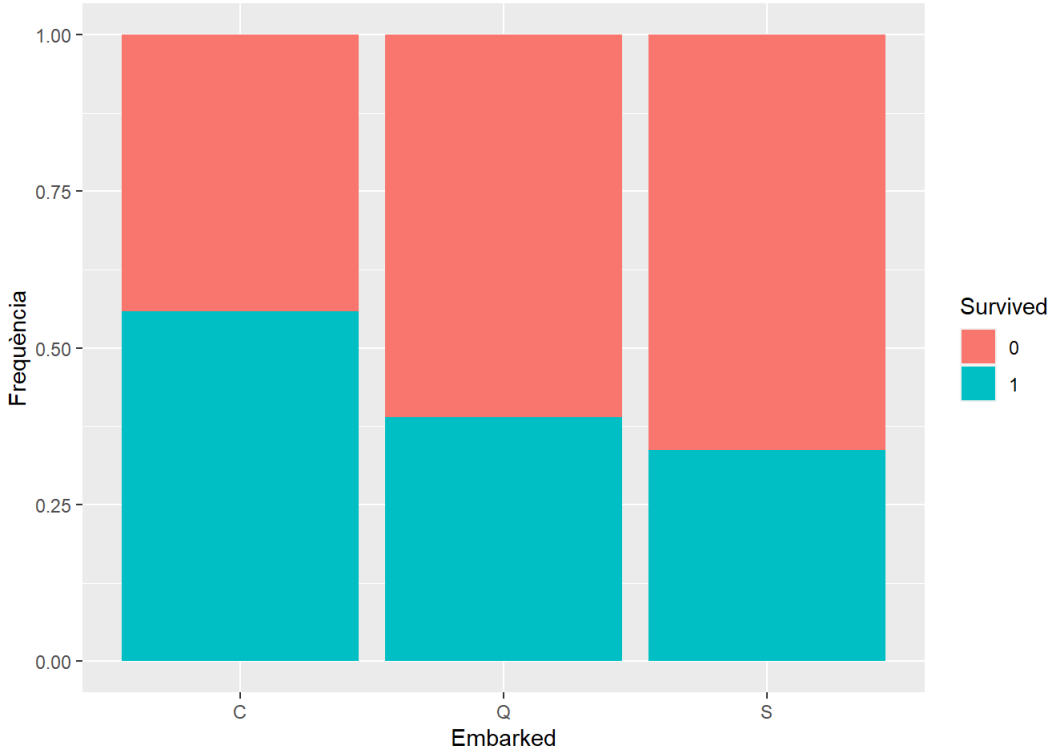
```
# Numero de mujeres en la embarcación
nrow(dtTitanic[dtTitanic$Sex == "female",])
```

```
## [1] 314
```

Como podemos ver el porcentaje de supervivencia de mujeres es mal alto que el de hombres, donde damos a entender que el grupo de mujeres con un 74,2% se le dio más prioridad a la hora de evacuar en los botes salvavidas al contrario de los hombres que es mas bajo con el resto de porcentaje. Cabe decir que el grupo de hombres también era mas grande que el de mujeres como podemos observar.

Otra comparación que hacemos es la supervivencia en función del embarcamiento.

```
# Plot en función del embarcamiento
ggplot(data=dtTitanic, aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+ylab("Frequència")
```



Obtenemos otra vez una tabla con las frequencias de probabilidades sobrevivir sobre el embarcamiento.

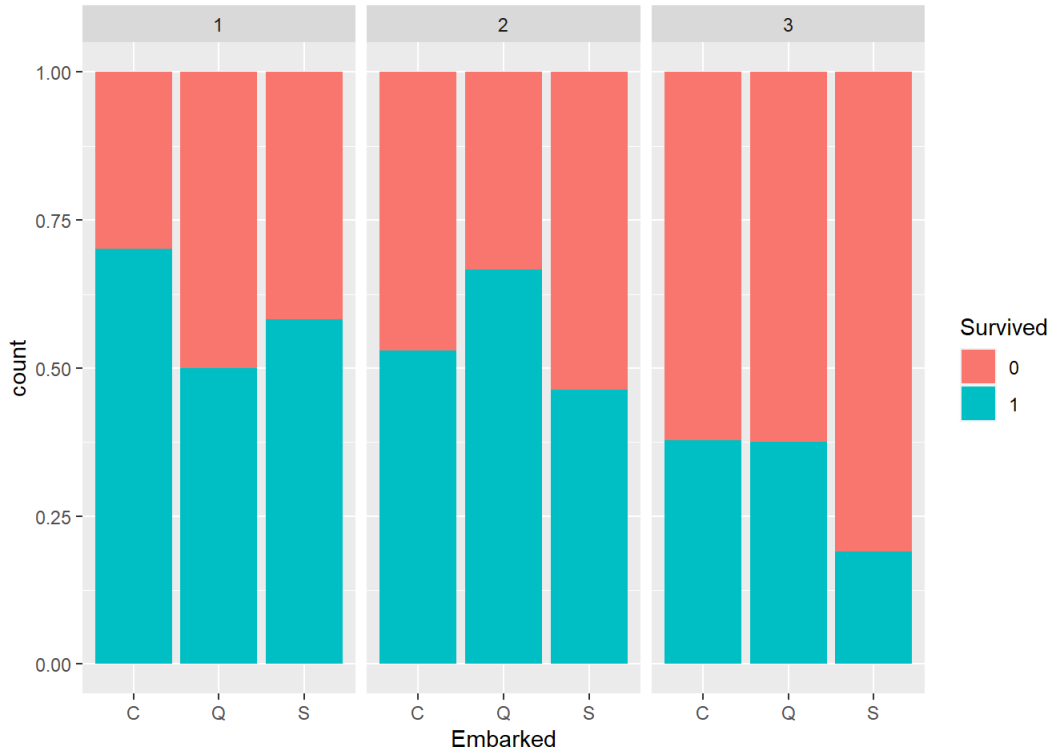
```
tabla_comparativa<-table(dtTitanic$Embarked,dtTitanic$Survived)
for (i in 1:dim(tabla_comparativa)[1]){
    tabla_comparativa[i,]<-tabla_comparativa[i,]/sum(tabla_comparativa[i,])*100
}
tabla_comparativa
```

```
##
##            0        1
##   C 44.11765 55.88235
##   Q 61.03896 38.96104
##   S 66.30435 33.69565
```

Donde vemos que el porcentaje de supervivencia mayor es en los pasajeros de embarcación C.

El siguiente gráfico que generaremos, sera uno de frequencias utilizando 3 variables, la variable objetivo de supervivencia, la utilizada Embarked, pero ademas le añadiremos el PClass, para ver en que clase viajaban los pasajeros.

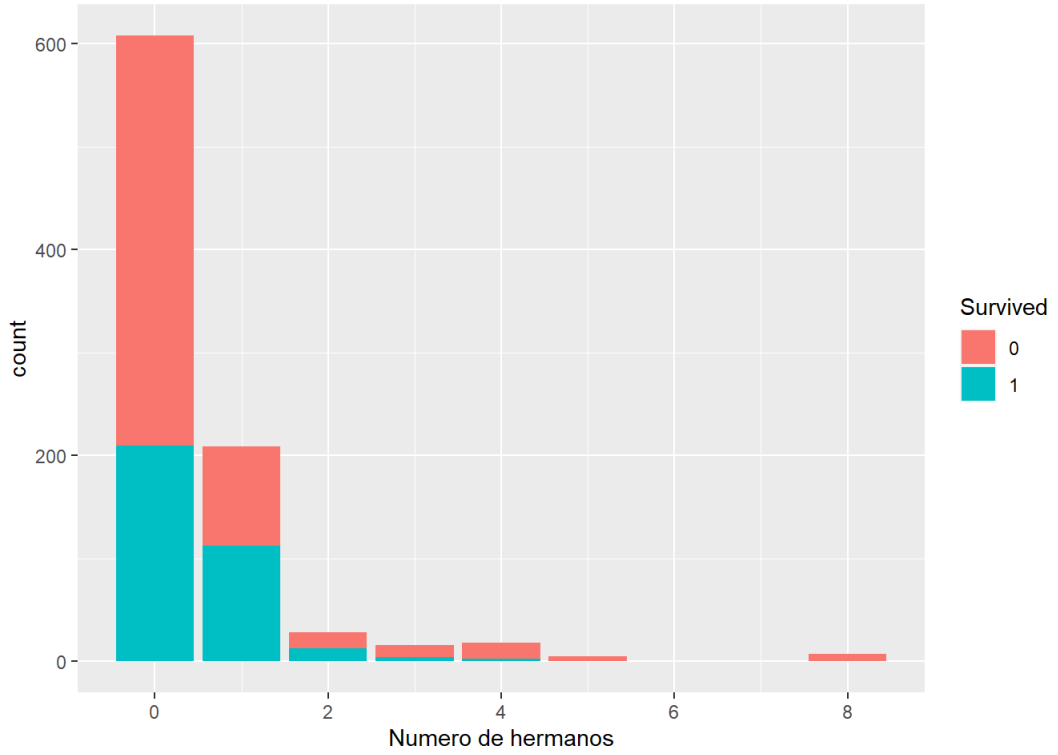
```
# Gráfico comprando embarked,survived y PcClass.
ggplot(data = dtTitanic,aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+facet_wrap(~Pclass)
```



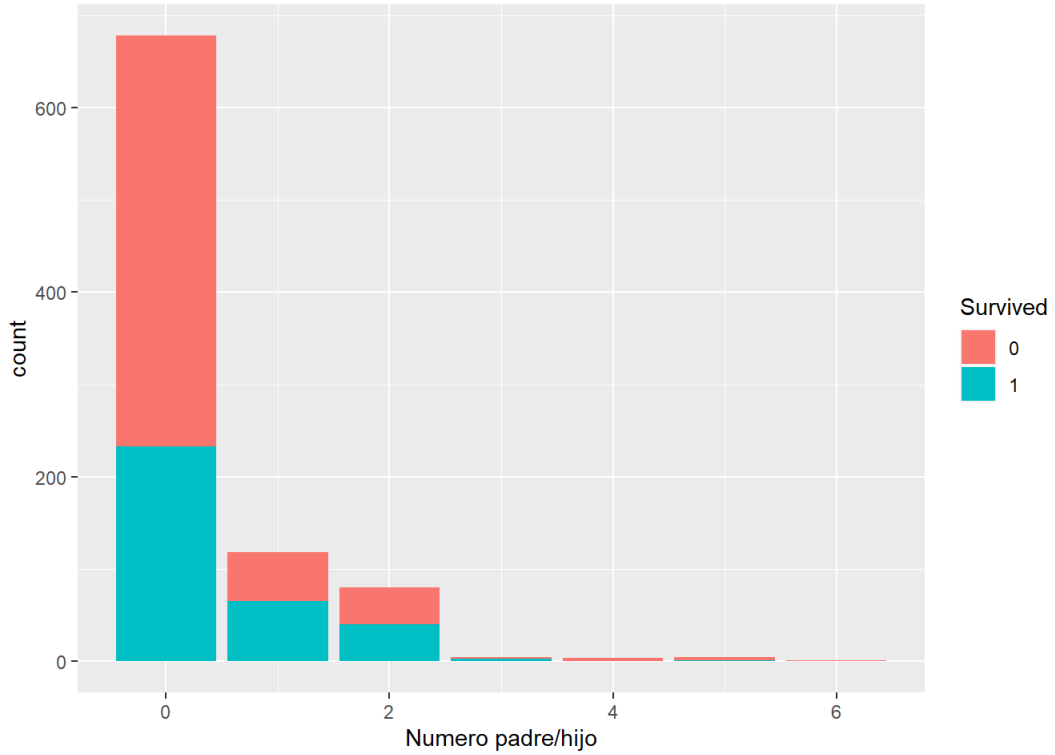
Podemos ver que los porcentajes son superiores en supervivencia si la clase es más alta.

En los siguientes gráficos de frequencias utilizaremos las variables SibSp y Parch junto a la variable objetivo Survived.

```
ggplot(data = dtTitanic,aes(x=SibSp,fill=Survived))+geom_bar()+xlab("Numero de hermanos")
```

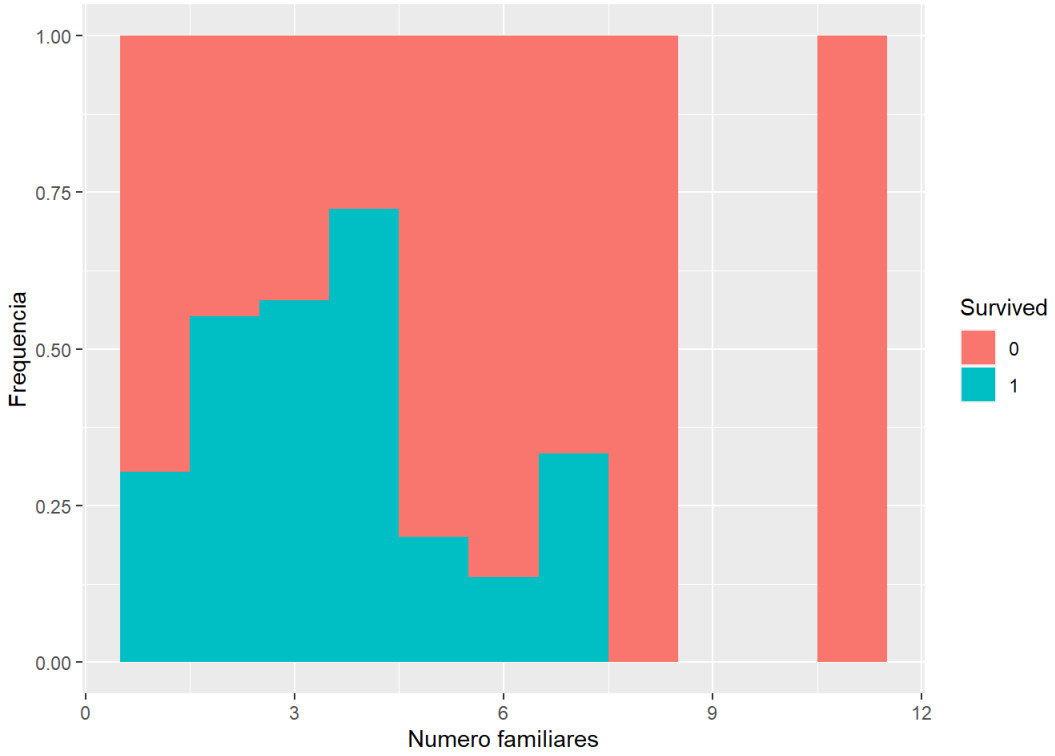



```
ggplot(data = dtTitanic,aes(x=Parch,fill=Survived))+geom_bar()+xlab("Numero padre/hijo")
```



Como podemos ver los gráficos son muy similares, por lo tanto indica presencia de correlaciones altas y por eso la creación de la variable nFamiliares que hemos hecho anteriormente y a partir de esta en el siguiente gráfico podremos ver cual ha sido la supervivencia en función del número de familiares.

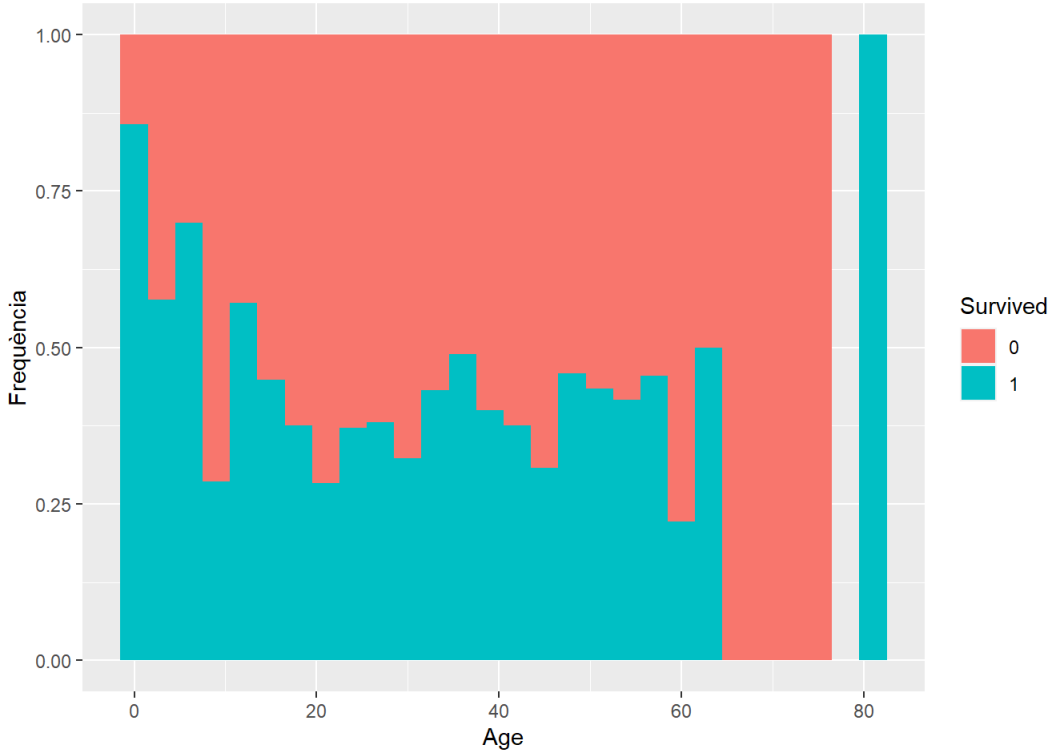
```
ggplot(data =
dtTitanic[!is.na(dtTitanic$nFamiliares),],aes(x=nFamiliares,fill=Survived))+geom_histogram(binwidth
=1,position="fill")+ylab("Frequencia")+xlab("Numero familiares")
```



Como podemos observar las familias menos numerosas son las que tienen el porcentaje de supervivencia más alto, viendo también que las familias mas altamente numerosas el porcentaje de supervivencia puede ser 0 y viendo que la frequencia en de 8 - 11 en numero de familiares es nula. El parametro position con fill podemos ver como nos muestra la acumulación de proporciones uno dentro de otro.

Con el gráfico anterior podemos ver que se pueden acumular muchas clases en un gráfico, ya hemos analizado la variable Age en otra variable factorizada funcion de si era adulto o niño, pero en el siguiente gráfico analizaremos en funcion de la edad númerica, utilizando un gráfico parecido al anterior, que nos acumula las proporciones y nos muestra las frequencias.

```
# Survival en función de la edad númerica
ggplot(data = dtTitanic[!is.na(dtTitanic$Age),],aes(x=Age,fill=Survived))+geom_histogram(binwidth = 3,position="fill")+ylab("Frequència")
```



Como podemos ver la frequencia de supervivencia va de más a menos en función va aumentando la edad, habiendo unos picos y bajones entre ellos, y viendo que al final que hubo supervivencia a alta edad.

Como se pide para la resolución de la práctica, se realizará un estudio sobre los datos utilizando analisis estadistico. Este estudio se realizará sobre la variable de edad, comparando tambien entre el sexo de los pasajeros. Con los ultimos gráficos se puede observar que la frequencia de supervivencia ronda entre los 40-60%, por lo tanto queremos ver si este porcentaje se mantiene si separamos la población por sexo.

Objetivo también es el verificar la normalidad, en nuestro caso con unos de los test más utilizados para eso. Son los tests de Saphiro-wilk y Kolmogrov-smirnov. Los dos comparan la distribución de los datos en una distribución normal. La hipotesis nula de nuestro analisis estadistico, es que la población esta distribuida con normalidad, si obtenemos el p-valor menor al nivel de significación que se usa generalmente,α=0,05, entonces dicha hipotesis es rechazada y se concluye que no hay una distribución normal. En caso contrario, si p-valor es mayor a nuestro nivel de significación, no se podra rechazar dicha hipotesis y asumiremos que tenemos una distribucion normal.

Mediante las funciones en R, ks.test() y shapiro.test(), representaremos dichos test para analizar la normalidad, ademas usaremos el metodo ecdf para ver de manera grafica la distribución comparando la edad de nuestros grupos de hombres y mujeres en nuestro conjunto de datos.

```
edad_hombre <- dtTitanic$Age[dtTitanic$Sex=="male"]
edad_mujer <- dtTitanic$Age[dtTitanic$Sex=="female"]


#Test de Kolmorgov-smirnov
ks.test(edad_mujer,edad_hombre)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  edad_mujer and edad_hombre
## D = 0.095315, p-value = 0.04971
## alternative hypothesis: two-sided
```

```
#Test de Saphiro-wilk
shapiro.test(dtTitanic$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dtTitanic$Age
## W = 0.95882, p-value = 3.969e-15
```

```
var.test(edad_mujer,edad_hombre)
```

```
##
##  F test to compare two variances
##
## data:  edad_mujer and edad_hombre
## F = 0.97982, num df = 313, denom df = 576, p-value = 0.8453
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8085848 1.1940983
## sample estimates:
## ratio of variances
##          0.9798199
```

Hemos podido ver en el gráfico que las dos lineas tienen una misma similitud en la forma, por lo tanto podemos ver la compensación en en edades en nuestro conjunto de datos. Pero viendo los resultados de los respectivos tests, vemos que el p-valor esta por debajo de el nivel de significación, por lo tanto rechazamos nuestra hipotesis de normalidad y podemos afirmar que no nuestro conjunto de datos no sigue una distribución normal.

A continuación realizaremos la prueba T student, una de las pruebas más utilizadas en la estadística. Teniendo como hipotesis nula que las medias de las edades en los grupos son iguales y teniendo como alternativas mayores y menores.

```
t.test(edad_mujer,edad_hombre)
```

```
##
##  Welch Two Sample t-test
##
## data:  edad_mujer and edad_hombre
## t = -2.5257, df = 648.52, p-value = 0.01179
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.0688028 -0.5093856
## sample estimates:
## mean of x mean of y
##  28.21673  30.50582
```

```
t.test(edad_mujer,edad_hombre,alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  edad_mujer and edad_hombre
## t = -2.5257, df = 648.52, p-value = 0.005893
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.7961705
## sample estimates:
## mean of x mean of y
##  28.21673  30.50582
```

```
t.test(edad_mujer,edad_hombre,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  edad_mujer and edad_hombre
## t = -2.5257, df = 648.52, p-value = 0.9941
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -3.782018       Inf
## sample estimates:
## mean of x mean of y
##  28.21673  30.50582
```

Como hemos podido ver en dos de los tests tenemos el p-valor por debajo del nivel significativo rechazando asi nuestra hipotesis, pero en la comprativa mayor el p-valor es superior.

Los datos tratados en esta práctica se guardaran limpios y tratados en un fichero adicional como se estipula en los puntos de la práctica.

```
write.csv(dtTitanic,"../DATASETS/clean-titanic.csv", row.names = FALSE)
```

2.5 Conclusion

Como todo documento y como se pide en la práctica, una vez realizado el estudio sobre nuestro conjunto de datos del titanic, establecemos una serie de conclusiones obtenidas sobre dicho estudio, que se ha centrado en ver la supervivencia de los pasajeros en la embarcación.

El conjunto de datos en si es un conjunto de datos, muy simple y limpio, y esto es debido a que es un conjunto de datos preparado para gente que se inicia en el mundo de la ciencia de datos. Se han visto valores nulos en la variable de edad, los cuales se han sustituido por la media de la edad de la embarcación. También se han utilizado diagramas de caja para encontrar esos valores atípicos en las variables de `Age`, `SibSp`, `Parch` i `Fare`, donde el valor más alto atipico ha sido en esta ultima variable y por eso se ha analizado posteriormente, pero también hemos podido determinar gracias a la estructuración y lo bien que esta la información en el conjunto de datos que los valores atípicos de las otras variables analizadas no son necesarios de eliminar por que pueden ser perfectamente útiles y se pueden considerar como valores tipicos para el analisis de los datos.

También hemos podido comprobar por diferentes variables, la frequencia y proporcion de supervivencia. La relación entre edades y entre edades clasificatorias de niño y adultos, donde hemos podido ver que los niños tenian una tasa de supervivencia mas alta que los adultos, y ademas también ver que en función de que aumentaba la edad la frequencia de supervivencia iba disminuyendo, viendo un valor atípico a alta edad de supervivencia. Al igual también hemos visto que la tasa de supervivencia de mujeres es superior a la de hombres, por lo tanto se puede ver que en ese momento la tripulación que gestionaba la evacuación en los botes salvavidas priorizaba niños y mujeres, viendo algun valor diferente como el de la persona de muy alta edad, que alomejor por clase social fue priorizada. Per el modelo de acierto mayor si fuera la pregunta seria que si era mujer o niño fue superviviente y si fue hombre no.

También como se pedia de objetivo en la práctica mediante los diferentes test aplicados, hemos podido llegar a la conclusión que las edades de nuestro conjunto de datos no seguia una distribución normal, en los grupos correspondientes de mujeres y hombres.

Como conclusión final, se propone como una resolución de una práctica muy visual y facil de entender para todas aquellas personas que quieran introducirse en este mundo, pudiendose así complicar mucho mas las cosas y utilizar millones de disposiciones gráficas para representar los datos de otras maneras y entrar mas en profundo desglosando más los datos.

2.6 Contribucion en la práctica

Contribucion	**Firma**
Investigacion previa	JFF
Desarrollo del documento	JFF
Desarollo del código	JFF