Práctica 2: Limpieza y validación de los datos

Alba Rollano Corroto

05 Enero 2021

Contents

1. Detalles de la activdad

1.1 Descripcion

En esta actividad se presenta un caso practico orientado a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integracion, limpieza, vvalidación y análisis de las mismas.

1.2 Objetivos

Los objetivos que se persiguen mediante el desarrollo de esta actividad práctica, son los siguientes:

- 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 validacion) 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 reultados para aportar conclusiones sobre el problema planteado en el proceso análitico
- Actuar con los principios éticos y legales relacionados con la manipulación de datos de datos en función del ámbito de aplicación
- Desarrollar las habilidades de aprendizaje que permita continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos

1.3 Competencias

Así, las competencias del Máster en Data Science que se desarrollan son:

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

2. Resolución

2.1 Descripcion del dataset

El 15 Abril de 1912 tuvo lugar uno de los momentos más destacados del S.XX, el hundimiento del Titanic, donde 1502 personas, entre pasajeros y personal, perdieron la vida.

Dada su calificación de el indestructible, no se consideró oprtuno la disponibilidad de tantas plazas en los botes salvavidas, como personas a bordo. Es por ello, que el conjuntos de datos de las personas salvadas y muertas hayan sido fruto de estudio en numerosas ocasiones, con el fin de extraer conclusiones sobre las prioridades establecidas a la hora de tomar el bote salvavidas o no.

En el presente estudio, se buscarán dichas conclusiones en base al conjunto de datos de Kaggle Titanic, divido en :

- Train.csv: dataset de entrenamiento con información de 891 personas
- Test.csv: dataset de validacion con in formación de 418 personas

Entre ambas muestras se recoge un total de 1309 registros de datos de los pasajeros y tripulación, evaluandose hasta 12 variables:

1. `PassengerId`: identificador numérico del pasajero
2. `Survived`: identificador de supervivencia (0= No sovrevivió, 1= sobrevivió) –> Sólo disponible en Train.csv
3. `PClass`: clase (1 = primera,2 = segunda, 3= tercera)
4. `Name`: nombre
5. `Sex`: sexo
6. `Age`: edad
7. `SibSp`: # de parientes abordo
8. `Parch`: # de padres/hijos a bordo
9. `Ticket`: número ticket
10. `Fare`: tarifa pasajero
11. `Cabin`: número cabina
12. `Embarked`: puerto de embarque (C= Cherbourg, Q= Queenstown, S= Southhampton)

2.2 Importancia y objetivos de los análisis

A partir del conjunto de datos train anteriormente presentado, se plantea el estudio sobre que factores influyeron de manera más relevante a la hora de determinar si un pasajero o tripulante podía subir al bote salvavidas, y por ende salvar su vida , o no. Además, en base a este dataset, se generará un modelo predictivo, que determine, en base a ciertas caracteristicas relevantes, si un pasajero se salvaría o no. Y el cuál se probará en el dataset test.csv

Adicionalmente, se podrá caracterizar a los pasajeros y tripulación que formaron parte de la embarcación Titanic, en lo que a distinción por genero, edad y clase se refiere.

2.3 Análitica descriptiva

En esta primera parte, se realizará un estudio inicial de los datos, así como la evaluación de las posibles relaciones existentes entre los diferentes parámetros.

2.3.1 Lectura del fichero

El primero de los pasos, será cargar los datos a evaluar:

```
# Carga de los datos test y train

test <- read.csv('C:/Users/Propietario/Desktop/Master en data science/Semestre 2/Tipologia y ciclo de v
train <- read.csv('C:/Users/Propietario/Desktop/Master en data science/Semestre 2/Tipologia y ciclo de

# Se unen los dos conjuntos de datos en uno solo
totalData <- bind_rows(train,test)
filas=dim(train)[1]

#Se muestra las primeras filas
head( totalData )
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                             Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                              Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                            Allen, Mr. William Henry   male  35     0
## 6                                    Moran, Mr. James   male  NA     0
##   Parch           Ticket    Fare Cabin Embarked
## 1     0        A/5 21171  7.2500              S
## 2     0         PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0           113803 53.1000  C123        S
## 5     0           373450  8.0500              S
## 6     0           330877  8.4583              Q
```

Y verificar la estructura de los mismos:

```
# Informacion de la estructura de los datos y sus variables
str(totalData)
```

```
## '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 (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

Donde se observa el tipo de dato asociado a cada variable.

2.3.1 Selección de los datos de interes

Dado que el objetivo es analizar la caracteristicas más relevantes a la hora de determinar la supervivencia de una persona, se considera que las columnas *Ticket*, *Fare* y *Cabina* carecen de influencia, y por ende, pueden ser eliminadas del dataset:

```
# Informacion de la estructura de los datos y sus variables
totalData <- totalData[, -c(9,10,11)]
```

No obstante, cabe destacar que la variable *Fare* podría ser sinificativa si quisieramos evaluar el precio medio del ticket en total o en base a ciertas características. Como para esta actividad se descarta dicho cálculo se puede eliminar.

Adicionalmente, las variables *PassengerID* y *Name* no aportan valor a la hora de extraer conclusiones, pero serán relevantes para identificar a cada pasajero o tripulante.

2.3.2 Elementos nulos y vacios

Uno de los aspectos más importantes en esta fase, es detectar la presencia de valores nulos o desconocidos. Normalmente, estos se deben a datos perdidos o desconocidos en el momento de recopilar los datos.

Por este motivo, en primera instancia, se deberán localizar:

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

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
##       SibSp       Parch    Embarked
##           0           0           0
```

```
colSums(totalData=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0          NA
##       SibSp       Parch    Embarked
##           0           0           2
```

Es decir, encontramos tanto valores NA como vacios, dentro de las variables *Survived*, *Age* y *Embarked*.

Llegados a este punto, se debe decidir como manejar estos registros. Una alternativas sería eliminar dichos datos, pero tendría una pérdida de información asociada. Los 418 registros desconocidos de la variable *Survived* corresponden con el dataset test.csv , y cuyo valor se ha de predecir a posteriori, por lo que se obviaran en este paso. Para *Age* y *Embarked* se procederá como sigue:

```
# Tomamos valor "C" para los valores vacíos de la variable "Embarked"
totalData$Embarked[totalData$Embarked==""]="C"

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

# Y Discretizamos
cols<-c("Survived","Pclass","Sex","Embarked")
for (i in cols){
  totalData[,i] <- as.factor(totalData[,i])
}
```

2.3.3 Valores extremos

Se entiende por valor extremo o *outlier*, aquellos datos que parecen no ser congruentes si los comparamos con el resto de los datos. Existen dos vías para identificar dichos valores dentro de un dataset:

- Representar un diagrama de caja por cada variable y evaluar si existe algún valor que dista mucho del rango intercuartílico (la caja)
- Mediante la función box.plot.stats() de R

Dada la simplicidad del mismo, se hará uso del segundo método para las variables *Age*, *Sibsp*, *Parch*. El resto de variables se obviaran por no ser numéricas o discretas:

```
boxplot.stats(totalData$Age)$out
```

```
##   [1]  2.00 58.00 55.00  2.00 66.00 65.00  0.83 59.00 71.00 70.50  2.00
##  [12] 55.50  1.00 61.00  1.00 56.00  1.00 58.00  2.00 59.00 62.00 58.00
##  [23] 63.00 65.00  2.00  0.92 61.00  2.00 60.00  1.00  1.00 64.00 65.00
##  [34] 56.00  0.75  2.00 63.00 58.00 55.00 71.00  2.00 64.00 62.00 62.00
##  [45] 60.00 61.00 57.00 80.00  2.00  0.75 56.00 58.00 70.00 60.00 60.00
##  [56] 70.00  0.67 57.00  1.00  0.42  2.00  1.00 62.00  0.83 74.00 56.00
##  [67] 62.00 63.00 55.00 60.00 60.00 55.00 67.00  2.00 76.00 63.00  1.00
##  [78] 61.00 60.50 64.00 61.00  0.33 60.00 57.00 64.00 55.00  0.92  1.00
##  [89]  0.75  2.00  1.00 64.00  0.83 55.00 55.00 57.00 58.00  0.17 59.00
## [100] 55.00 57.00
```

```
boxplot.stats(totalData$SibSp)$out
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3
## [36] 5 4 3 4 8 4 3 4 8 4 8 3 4 5 3 4 8 4 8 4 3 3
```

```
boxplot.stats(totalData$Parch)$out
```

```
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1
##  [36] 2 1 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1
##  [71] 1 2 1 2 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2
## [106] 2 3 4 1 2 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1
## [141] 2 1 1 2 5 2 1 1 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1
## [176] 1 2 1 2 3 1 2 1 2 2 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1
## [211] 1 5 2 1 1 1 1 3 1 2 2 1 2 1 2 1 2 4 1 1 2 1 1 1 4 6 2 3 1 1 2 2 2 1 1
## [246] 2 5 2 3 2 1 1 1 2 1 2 2 2 1 2 1 1 2 1 2 1 2 1 2 2 1 1 1 1 1 2 1 1 2 1
## [281] 1 1 2 1 2 9 1 1 1 2 2 2 1 9 1 1 2 2 1 1 2 1 1 1 1 1 1
```

Si se analizan dichos valores, se observa que son coherentes y por tanto, no descartables. Por lo que no se realizará ninguna acción al respecto.

2.3.4 Exportación de datos

Una vez acometidos los procesos de integración, validación y limpieza sobre el dataset inicial, se procede a guardarlos bajo un nuevo fichero denominado Titanic_clean.csv:

```
# Exportación del dataset limpio en .csv

write.csv(totalData,'C:/Users/Propietario/Desktop/Master en data science/Semestre 2/Tipologia y ciclo d
```

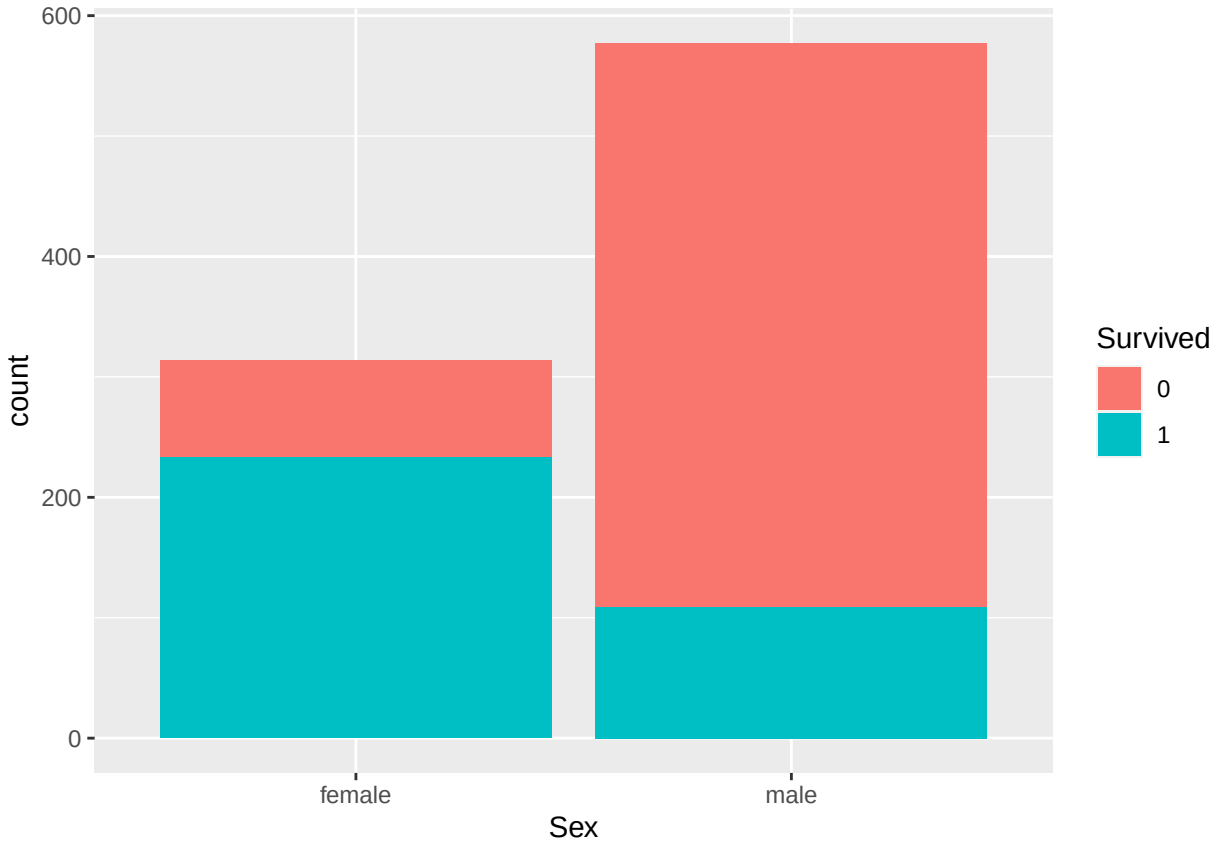
2.4 Análisis de los datos

En esta segunda fase de la actividad, tomando como referencia la base de datos trabajada, se trabajará en buscar relaciones entre las diferentes variables.

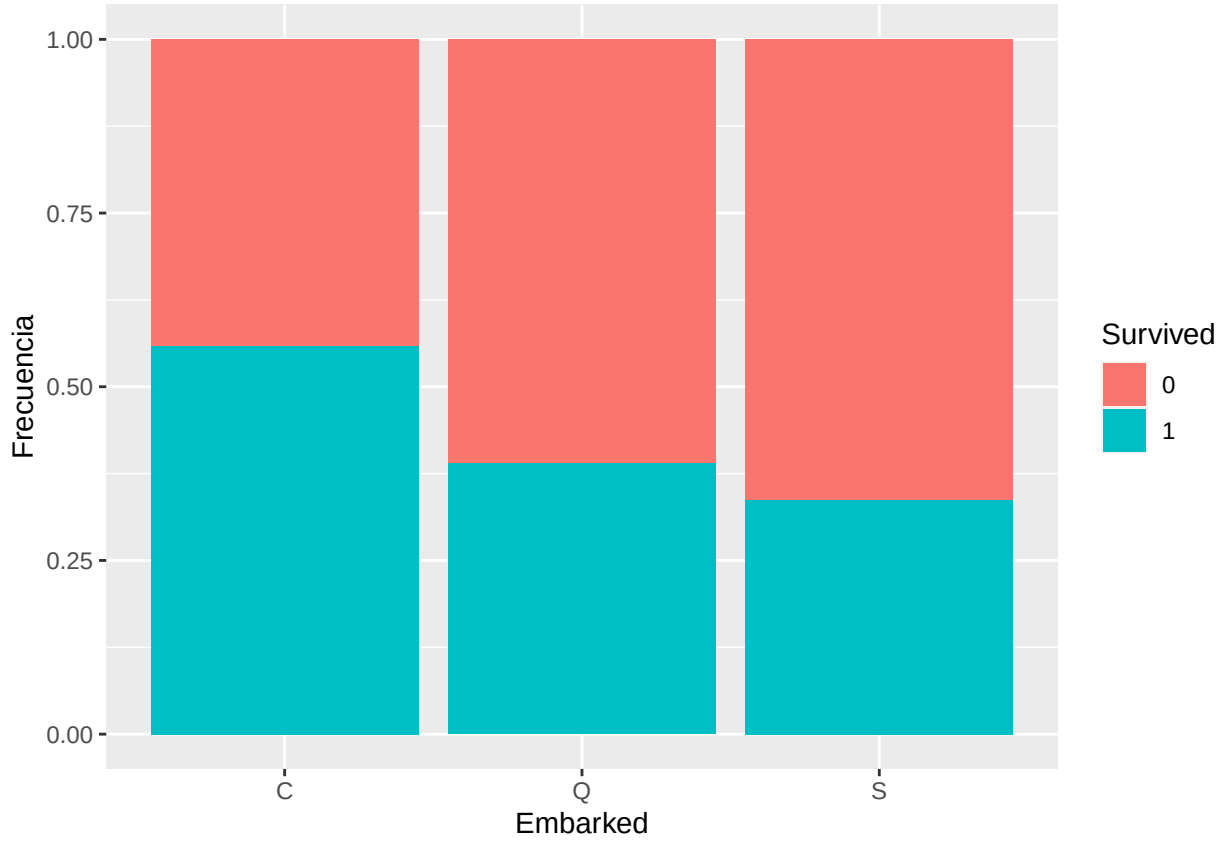
2.4.1 Visualización previa

El primer paso será representar gráficamente los atributos de la muestra con el fin de extraer información y conclusiones previas del conjunto de los datos de estudio:

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




```
# Otro punto de vista. Survival como función de Embarked:
ggplot(data = totalData[1:filas,],aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+ylab("Frecue
```



De donde se extrae:

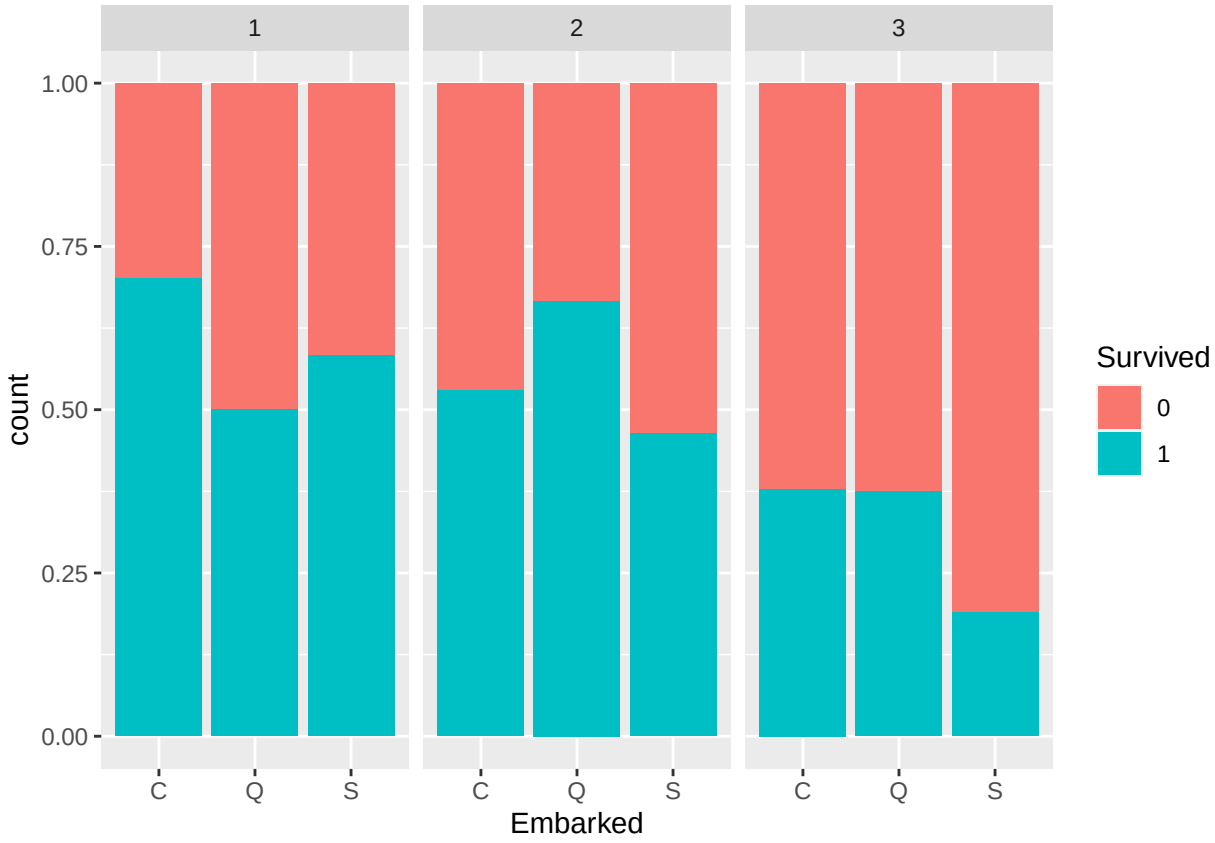
- El número de hombres que embarcaron eran casi el doble que el de mujeres
- Los supervivientes eran en su mayoria mujeres (casi el doble que de hombres)
- El mayor ratio de los supervivientes embarcaron en Cherbourgh, pudiendo obtener las tasas de supervivencia mediante el uso de la matriz de porcentages de frecuencia. Por ejemplo, la probabilidad de sobrevivir si se embarcó en "C" es de un 55,88%

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

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

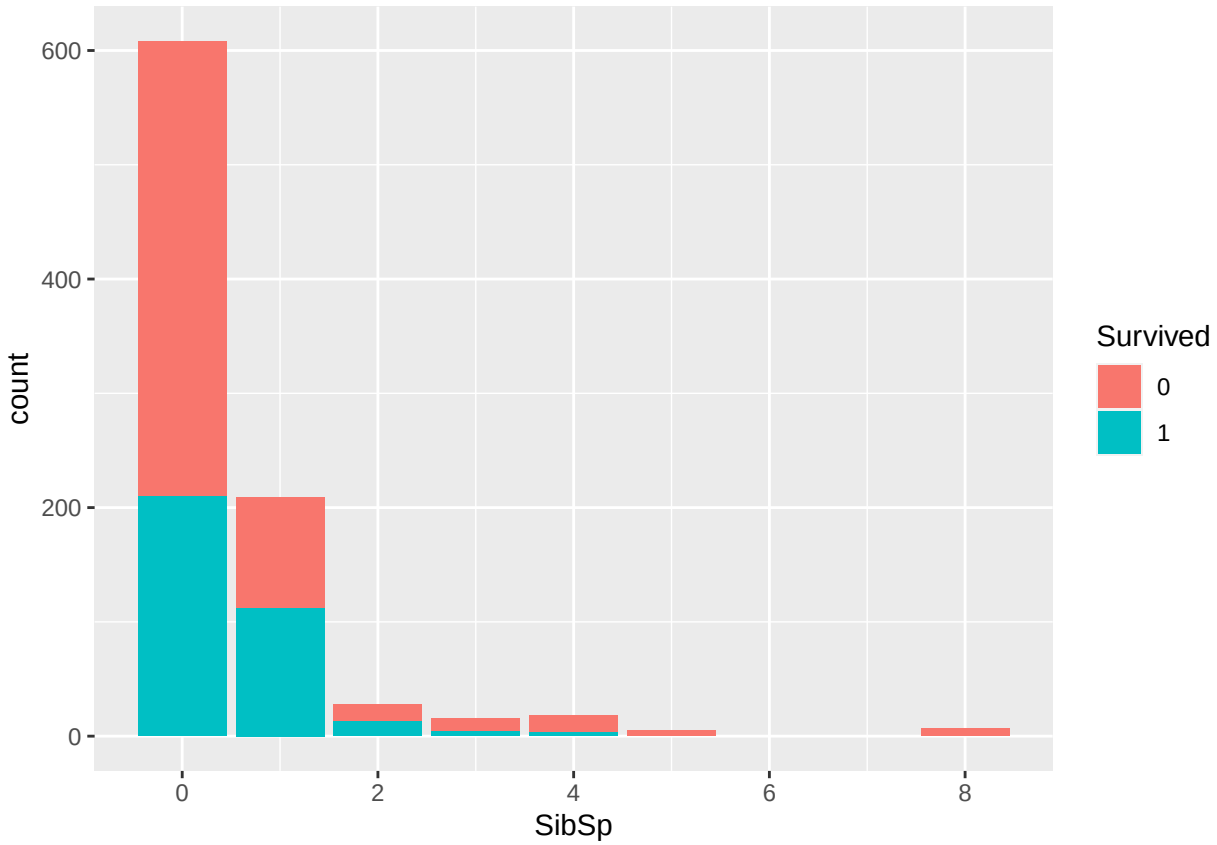
De cara a trabajar con las 3 variables: *Embarked*, *Survived* y *Pclass*, se puede obtener el gráfico de frecuencias:

```
# Ahora, podemos dividir el gráfico de Embarked por Pclass:
ggplot(data = totalData[1:filas,],aes(x=Embarked,fill=Survived))+geom_bar(position="fill")+facet_wrap(~
```

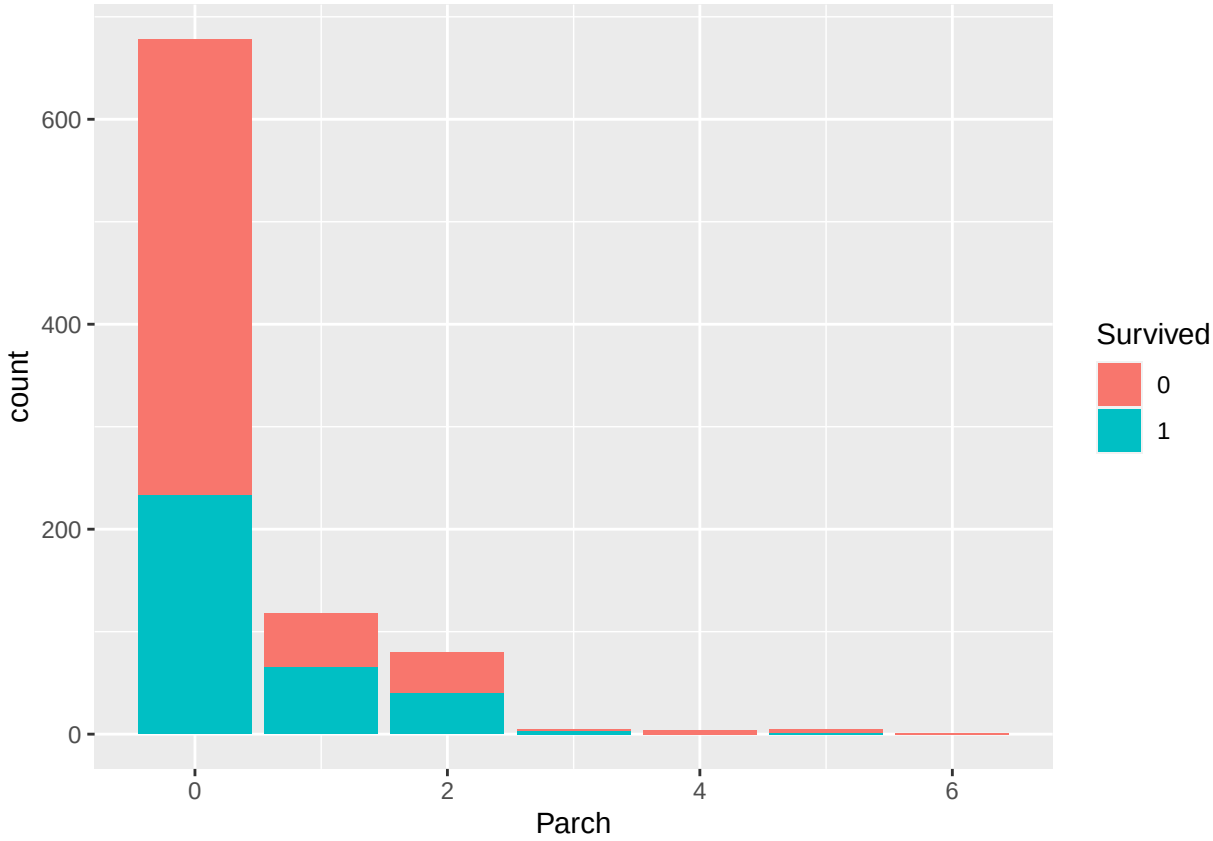


O comparando de dos en dos gráficos de frecuencias: Survived-SibSp y Survived-Parch:

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



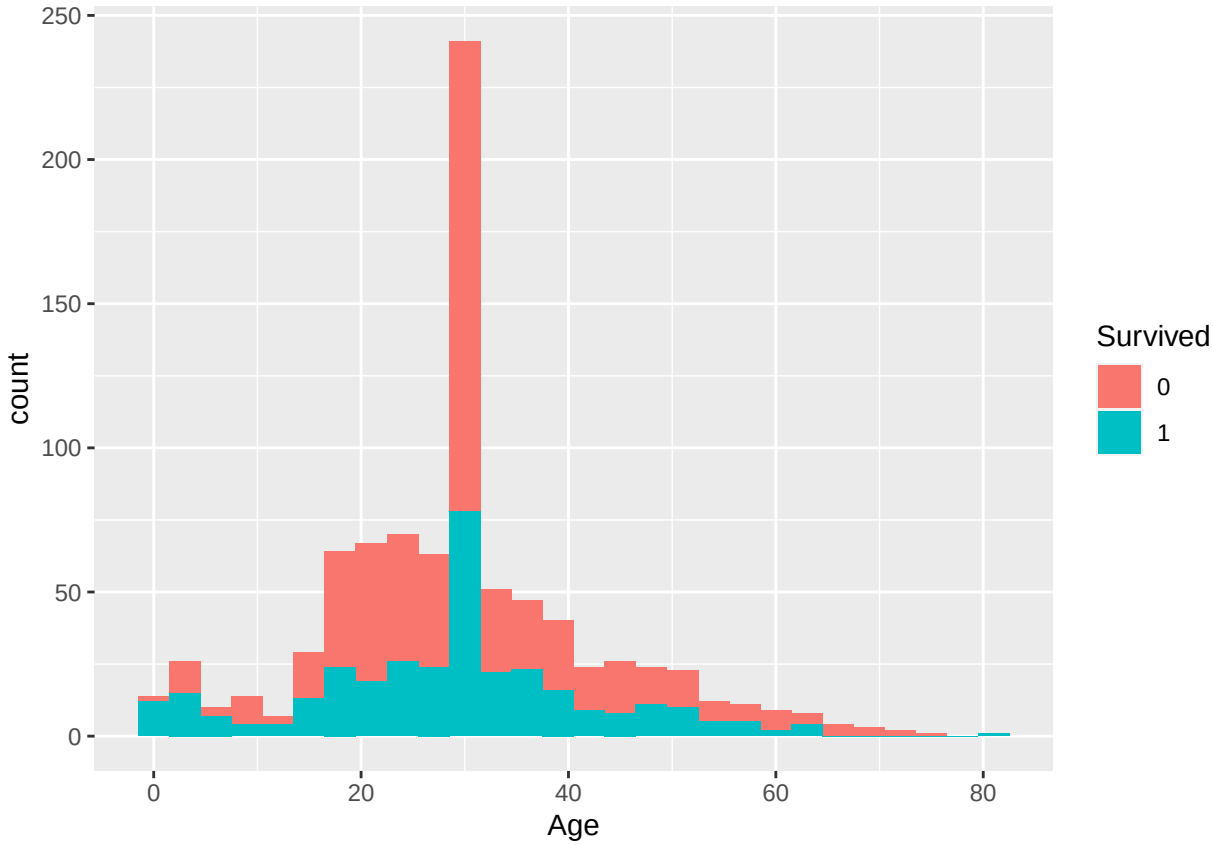

```
ggplot(data = totalData[1:filas,],aes(x=Parch,fill=Survived))+geom_bar()
```




```
# Se observa como las forma de estos dos gráficos es similar. Este hecho puede indicar presencia de cor
```

En última instancia, como overview, se pueden comparar los atributos Age y Survived, siendo el parámetro position="fill" la proporción acumulada de un atributo dentro de otro.

```
# Survival como función de age:
totalData1<-totalData[1:filas,]
ggplot(data = totalData1[!(is.na(totalData[1:filas,]$Age)),],aes(x=Age,fill=Survived))+geom_histogram(bi
```




```
ggplot(data = totalData1[!is.na(totalData[1:filas,]$Age),],aes(x=Age,fill=Survived))+geom_histogram(bin
```

Es decir, los mayores de ratio de supervivientes, se dan para los menores de 8 años y los mayores de 80.

De estos gráficos se obtiene información muy valiosa que se puede complementar con las tablas de contingencia (listadas abajo). Por un lado, la cantidad de pasajeros que sobrevivieron es menor en hombres y mujeres (hombres: 109 y mujeres 233). Si ademas se compará con la cantidad de hombres y mujeres que no sobrevivieron (81 mujeres y 468 hombres), la tasa de muerte en hombres es muchísimo mayor (el 81% de los hombres murieron mientras que en mujeres ese porcentaje baja a 25,8%).

En cuanto a la clase en la que viajaban, los pasajeros que viajaban en primera clase fueron los únicos que el porcentaje de supervivencia era mayor que el de mortalidad. El 62,96% de los viajeros de primera clase sobrevivió, el 47,2% de los que viajaban en segunda clase mientras que de los viajeros de tercera y de la tripulación solo sobrevivieron un 24,23% respectivamente. Para finalizar, destacamos que la presencia de pasajeros adultos era mucho mayor que la de los niños (2092 frente a 109).

```
tabla_SST <- table(totalData$Sex, totalData$Survived)
tabla_SST
```

```
##
##            0   1
##   female  81 233
##   male   468 109
```

```
prop.table(tabla_SST, margin = 1)
```

```
##
##                  0         1
```

```
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

```
tabla_SST <- table(totalData$Pclass, totalData$Survived)
tabla_SST
```

```
##
##       0   1
##   1  80 136
##   2  97  87
##   3 372 119
```

```
prop.table(tabla_SST, margin = 1)
```

```
##
##             0         1
##   1 0.3703704 0.6296296
##   2 0.5271739 0.4728261
##   3 0.7576375 0.2423625
```

2.4.2 Selección de los grupos de datos a analizar

A continuación, se seleccionan los grupos dentro del conjunto de datos que pueden resultar de interes analizar y comparar:

```
#Agrupación por genero
totalData.male <- totalData[totalData$Sex == 'male',]
totalData.female <- totalData[totalData$Sex == "female",]

#Agrupación por clase
totalData.primera <- totalData[totalData$Pclass == 1,]
totalData.segunda <- totalData[totalData$Pclass == 2,]
totalData.tercera <- totalData[totalData$Pclass == 3,]

#Agrupación por puerto de embarque
totalData.S <- totalData[totalData$Embarked == "S",]
totalData.C <- totalData[totalData$Embarked== "C",]
totalData.Q <- totalData[totalData$Embarked== "Q",]

#Agrupación por supervivencia
totalData.superviviente <- totalData[totalData$Survived == 1,]
totalData.nosuperviviente <- totalData[totalData$Survived == 0,]
```

2.4.3 Comprobación de la normalidad de la varianza

Según el teorema del límite central, para muestras mayores a 30 registros se puede asumir que sigue una distribución normal, no obstante, se verifica dicha normalidad por atributo en base a la prueba de normalidad *Anderson-Darling*.

Así, se comprobará para cada prueba el p-valor. Si dicho valor es mayor al nivel de significación $\alpha = 0,05$, se puede considerar que dicha variable en cuestión sigue una distribución normal.

```
library(nortest)
```

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

```
alpha = 0.05
col.names = colnames(totalData)

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

      # Salida por pantalla
      if (i < ncol(totalData) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Las Variables que no siguen una distribución normal son:
## PassengerId, Age,
## SibSp, Parch
```

2.4.4 Pruebas estadísticas

Como evaluación de la muestra, se empleará el estadistico Chi Cuadrado para validar la hipótesis nula que la proporción de hombres que han sobrevivido es mayor a la de mujeres:

```
mytable <- xtabs(~   Sex + Survived, data = totalData)
addmargins(mytable)
```

```
##         Survived
## Sex        0   1 Sum
##   female  81 233 314
##   male   468 109 577
##   Sum    549 342 891
```

```
#Se estima el estadístico chi cuadrado
chisq.test(mytable)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mytable
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

Dado que p-value $< 0{,}05$, se puede afirmar con un nivel de significacion de 0,05 que se rechaza la hipotesis nula.

2.4.5 Regresión

Por ultimo, como parte del estudio inferencial, se procederá a un análisis regresional multivariante. El objetivo es determinar la dependencia de la variable *Survived*, en función de las variables: *Sex*, *Age*,*Class*,*Sibsp*,*Parch*,*Embarked*.

A continuación se construye el modelo de regresión:

```
#Model= lm(formula = Survived ~ Sex, data = totalData[0:891,])
#summary(Model)
```

3.Conclusiones

Como se ha visto, se han realizado tanto análisis estádisticos como gráficos a partir de los cuales se han podido extraer las siguientes conclusiones: - El ratio de supervivientes mujeres es mayor al de los hombres - El porcentaje de hombres que embarcaron era mayor al de mujeres - Los pasajeros de 1 clase tenían mayor probabilidad de tomar un bote salva vidas - Niños y mayores presentaron mayor ratio de supervivencia

Previamente, se han sometido los datos a un preprocesamiento para manejar los casos de ceros o elementos vacíos y valores extremos (outliers). Para el caso del primero, se ha hecho uso de un método de imputación de valores de tal forma que no se deba eliminar registros del conjunto de datos inicial y que la ausencia de valores no implique llegar a resultados poco certeros en los análisis. Para el caso del segundo, el cual constituye un punto delicado a tratar, se ha optado por incluir los valores extremos en los análisis dado que parecen no resultar del todo atípicos si los comparamos con los valores que toman las correspondientes variables .

4.Bibliografia

1. Dalgaard, P. (2008). Introductory statistics with R. Springer Science & Business Media.
2. Vegas, E. (2017). Preprocesamiento de datos. Material UOC.
3. Gibergans, J. (2017). Regresión lineal múltiple. Material UOC.
4. Rovira, C. (2008). Contraste de hipótesis. Material UOC.