PRA2_Tipologia y Ciclo de vida de los datos

Ramon Martinez

1/3/2019

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

He escogido el juego de datos de kaggle Titanic. https://www.kaggle.com/c/titanic/data

El dataset describe los pasajeros que iban a bordo del RMS Titanic y que protagonizó uno de los hundimientos más famosos de la historia el 15 de Abril de 1912. 1502 personas de un total de 2224 perdieron la vida, uno de los motivos principales del gran número de pérdidas fue debido a la falta de suficientes salvavidas para los pasajeros y la tripulación.

Tal y como describe la web de kaggle, a pesar de que la suerte estaba involucrada en las posibilidades de supervivencia o hundimiento, algunos grupos de personas tenían más opciones que otros, tales como las mujeres, niños o clases superiores.

La finalidad de este dataset es la de analizar y predecir que pasajeros tenían más probabilidades de sobrevivir.

2. Integración y selección de los datos de interés a analizar.

El dataset está compuesto de 891 entradas (personas) con 12 atributos diferentes:

PassangerId, que identifica al pasajero, es un integer Survived, si sobrevivió o no Pclass, Clase de viaje Name, Nombre del pasajero Sex, Género Age, Edad del pasajero SibSp, Número de hermanos/esposa a bordo Parch, número de padre/hijo a bordo Ticket Fare Cabin Embarked, Puerto de embarque, C-Cherbourg, S-Southampton, Q-Queenstown

Como vemos, disponemos de atributos de los que podemos prescindir para nuestro análisis principal que será el de intentar predecir que pasajeros estaban más predispuestos a sobrevivir o no dependiendo de su sexo, edad y clase en la que viajaban. Analizaremos el atributo Survived que nos dirá si sobrevivió o no, la clase en la que viajaba, el sexo, la edad, si tenían hermanos/cónyuge o hijos/padres.

```
#cargamos el dataset en la variable titanic
titanic <- read.csv(file="/Users/ramon/pra2_tipologia/titanic.csv", header = TRUE)
#dimensión del dataset
dim(titanic)
```

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

```
#Sumarizamos la información que tenemos
summary(titanic)
```

```
##   PassengerId       Survived          Pclass     
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000  
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000  
##  Median :446.0   Median :0.0000   Median :3.000  
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309  
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000  
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000  
##                                                  
##                                     Name         Sex           Age       
##  Abbing, Mr. Anthony                   :  1   female:314   Min.   : 0.42  
##  Abbott, Mr. Rossmore Edward           :  1   male  :577   1st Qu.:20.12  
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1                Median :28.00  
##  Abelson, Mr. Samuel                   :  1                Mean   :29.70  
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00  
```

```
##  Adahl, Mr. Mauritz Nils Martin       :  1               Max.   :80.00
##  (Other)                              :885               NA's   :177
##      SibSp           Parch             Ticket         Fare
##  Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
##  Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
##                                   (Other) :852
##          Cabin     Embarked
##             :687    :  2
##  B96 B98    :  4   C:168
##  C23 C25 C27:  4   Q: 77
##  G6         :  4   S:644
##  C22 C26    :  3
##  D          :  3
##  (Other)    :186
```

```
#Eliminamos las columnas que no nos interesan
titanic <- titanic[, -(1)]
titanic <- titanic[, -(3)]
titanic <- titanic[, -(7:11)]
```

3. Limpieza de los datos.

Comprobamos los tipos de datos de los que disponemos. Como se aprecia, hay atributos que debemos cambiar. Survived, Pclass, es un valor discreto, por lo tanto lo cambiamos a factor.

```
#Comprobamos que tipo de datos tenemos
sapply(titanic, class)
```

```
##  Survived    Pclass       Sex       Age     SibSp     Parch
## "integer" "integer"  "factor" "numeric" "integer" "integer"
```

```
#Vemos que atributos son los subceptibles de ser factorizados, aquellas con pocas clases. Es decir los
apply(titanic,2, function(x) length(unique(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch
##        2        3        2       89        7        7
```

```
#convertimos los valores
titanic$Survived <- as.factor(titanic$Survived)
titanic$Pclass <- as.factor(titanic$Pclass)
#Comprobamos de nuevo tras convertirlos
sapply(titanic, class)
```

```
##  Survived    Pclass       Sex       Age     SibSp     Parch
##  "factor"  "factor"  "factor" "numeric" "integer" "integer"
```

```
head(titanic)
```

```
##   Survived Pclass    Sex Age SibSp Parch
## 1        0      3   male  22     1     0
## 2        1      1 female  38     1     0
## 3        1      3 female  26     0     0
## 4        1      1 female  35     1     0
## 5        0      3   male  35     0     0
## 6        0      3   male  NA     0     0
```

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Vemos que tenemos 177 entradas sin la edad, tenemos diferentes opciones para gestionar estos valores NA (Not Available):

- Eliminar directamente las entradas que contengan estos valores.
- eliminar la columna, que en nuestro caso no es buena idea ya que la edad es relevante.
- Imputar los valores perdidos, a través del cálculo de la mediana, media, kNN, etc.
- Convertir a categórico y filtrar los valores NaN en una categoría propia.

Si eliminamos las entradas que no contienen la edad perdemos bastantes datos, por lo que en principio vamos a evitarlo y nos centraremos en reemplazar esos valores por la media ya que no tenemos grandes outliners.

```
sapply(titanic, function(x) sum(is.na(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch
##        0        0        0      177        0        0
```

La media es de 29.6, como todas las edades a excepción de los bebés menores de 1 año no tienen parte decimal vamos a redondear al alza para sustituir los valores NA haciendo uso de ceiling. Es decir, nos quedaría en 30.

```
ceiling(mean(titanic$Age, na.rm=T))
```

```
## [1] 30
```

```
titanic$Age[is.na(titanic$Age)] <- ceiling(mean(titanic$Age, na.rm=T))
```

Hemos sustituido los valores NA (Not Available) por la media de los valores de la edad quitando los NA para su cálculo.

```
sapply(titanic, function(x) sum(is.na(x)))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch
##        0        0        0        0        0        0
```

Detección de campos vacios

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

```
## Survived   Pclass      Sex      Age    SibSp    Parch
##        0        0        0        0        0        0
```

3.2. Identificación y tratamiento de valores extremos.

La identificación de los valores extremos es importante ya que puede producir cambios, tendencias en la exactitud de nuestras predicciones y estimaciones. Por lo tanto es esencial saber si debemos dejarlos o no.

Los atributos categóricos tan solo nos aseguraremos de que no contienen valores no deseados, el único valor numérico al que podremos aplicar el boxplot es el de la edad y está representado abajo.

Como vemos, Survived solo tiene dos niveles, por lo tanto no tendríamos valores extremos.

```
table(titanic$Survived)
```

```
##
##   0   1
## 549 342
```

En la clase, tenemos primera, segunda y tercera clase, por lo tanto tampoco identificamos valores anomalos o extremos.

```
table(titanic$Pclass)
```

```
##
```

```
##   1   2   3
## 216 184 491
```

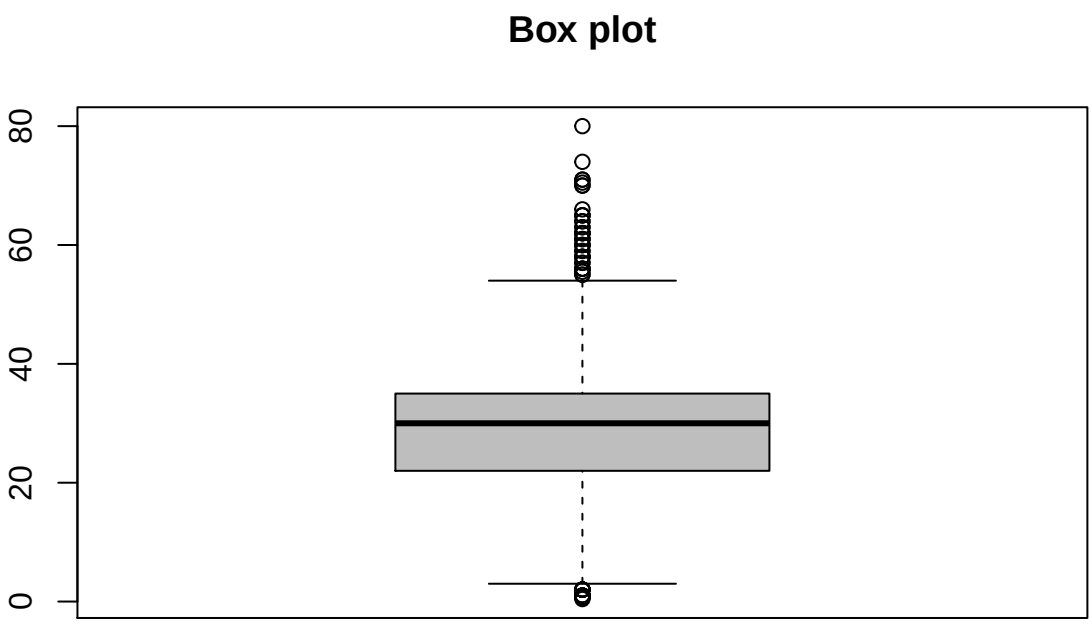
Lo mismo sucede con el género, female o male.

```
table(titanic$Sex)
```

```
##
## female   male
##    314    577
```

En el valor continuo edad, sí que detectamos valores extremos como se puede apreciar en la gráfica inferior, estos valores están entre los 66 y los 80 años por lo que a pesar de que hay pocas personas en esta edad son valores normales y que tiene sentido dejar.

```
boxplot(titanic$Age,main="Box plot", col="gray")
```

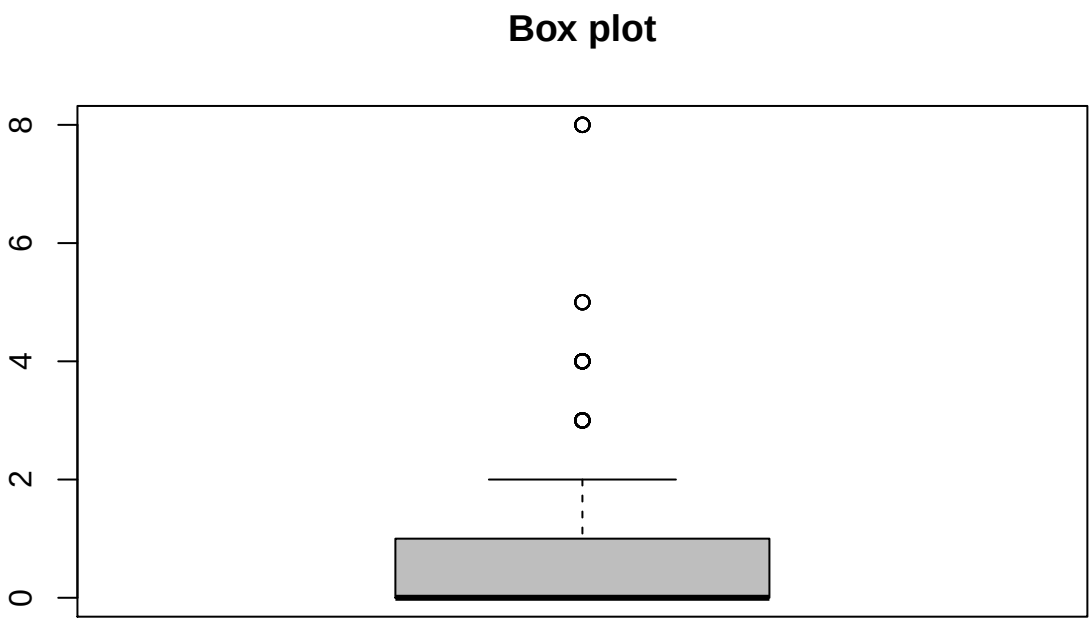



```
boxplot.stats(titanic$Age)
```

```
## $stats
## [1]  3 22 30 35 54
##
## $n
## [1] 891
##
## $conf
## [1] 29.31188 30.68812
##
## $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
```

Tenemos algunos valores extremos que van hasta 8 hermanos a bordo, aunque es un valor elevado puede ser perfectamente correcto por lo tanto los dejamos.

```
boxplot(titanic$SibSp,main="Box plot", col="gray")
```

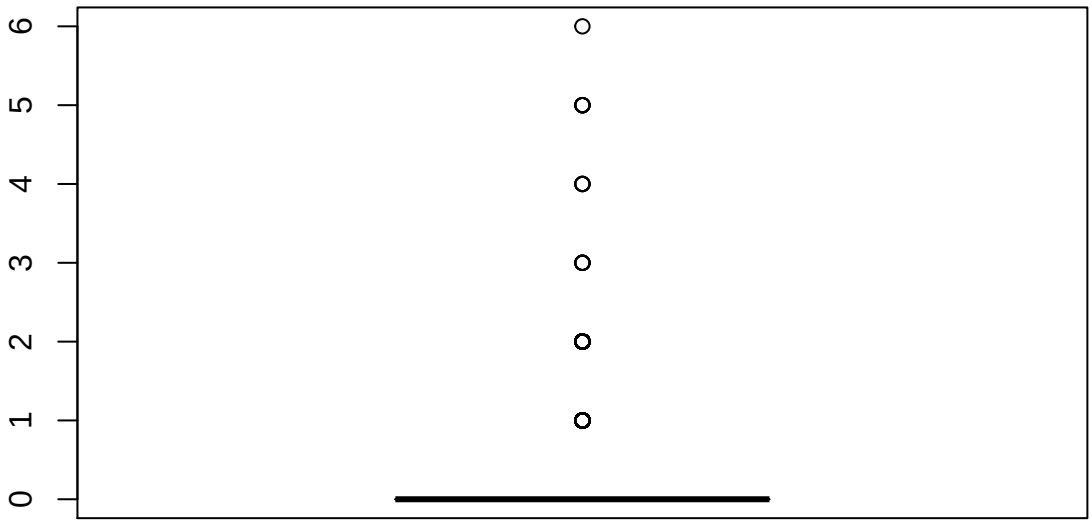



```
boxplot.stats(titanic$SibSp)
```

```
## $stats
## [1] 0 0 0 1 2
##
## $n
## [1] 891
##
## $conf
## [1] -0.05293199  0.05293199
##
## $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
```

Así mismo, tenemos valores extremos en el número de hijos que llega hasta los 6, al ser un valor posible lo dejamos también.

```
boxplot(titanic$Parch,main="Box plot", col="gray")
```




```
boxplot.stats(titanic$Parch)
```

```
## $stats
## [1] 0 0 0 0 0
##
## $n
## [1] 891
##
## $conf
## [1] 0 0
##
## $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
```

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

De los atributos seleccionados

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

```
library(nortest)
alpha = 0.05
col.names = colnames(titanic)
for (i in 1:ncol(titanic)) {
 if (i == 1) cat("Variables que no siguen una distribución normal:\n")
 if (is.integer(titanic[,i]) | is.numeric(titanic[,i])) {
  p_val = ad.test(titanic[,i])$p.value
  if (p_val < alpha) {
   cat(col.names[i])
```

```
# Format output
   if (i < ncol(titanic) - 1) cat(", ")
   if (i %% 3 == 0) cat("\n")
  }
 }
}
```

```
## Variables que no siguen una distribución normal:
## Age, SibSpParch
```

Para ver si las variables están normalizadas aplicamos el test shapiro en cada variable numérica.

```
shapiro.test(titanic$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Age
## W = 0.95943, p-value = 5.304e-15
```

```
shapiro.test(titanic$SibSp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$SibSp
## W = 0.51297, p-value < 2.2e-16
```

```
shapiro.test(titanic$Parch)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Parch
## W = 0.53281, p-value < 2.2e-16
```

El resultado nos indica que ninguna variable está normalizada ya que el p-value es menor que el coficiente 0.05.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

No podemos usar métodos de correlación debido a que la variable Survived es categórica y no numérica.

5. Representación de los resultados a partir de tablas y gráficas.

Tabla con la relación de spervivientes, 342 sobrevivieron de los analizados en nuestro dataset, y 549 perecieron.

```
table(titanic$Survived)
```

```
##
##   0   1
## 549 342
```

Tabla con la relación de personas por clase de viaje.

```
table(titanic$Pclass)
```

```
##
##   1   2   3
## 216 184 491
```

Gráfica comparativa de supervivientes en función del sexo.

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

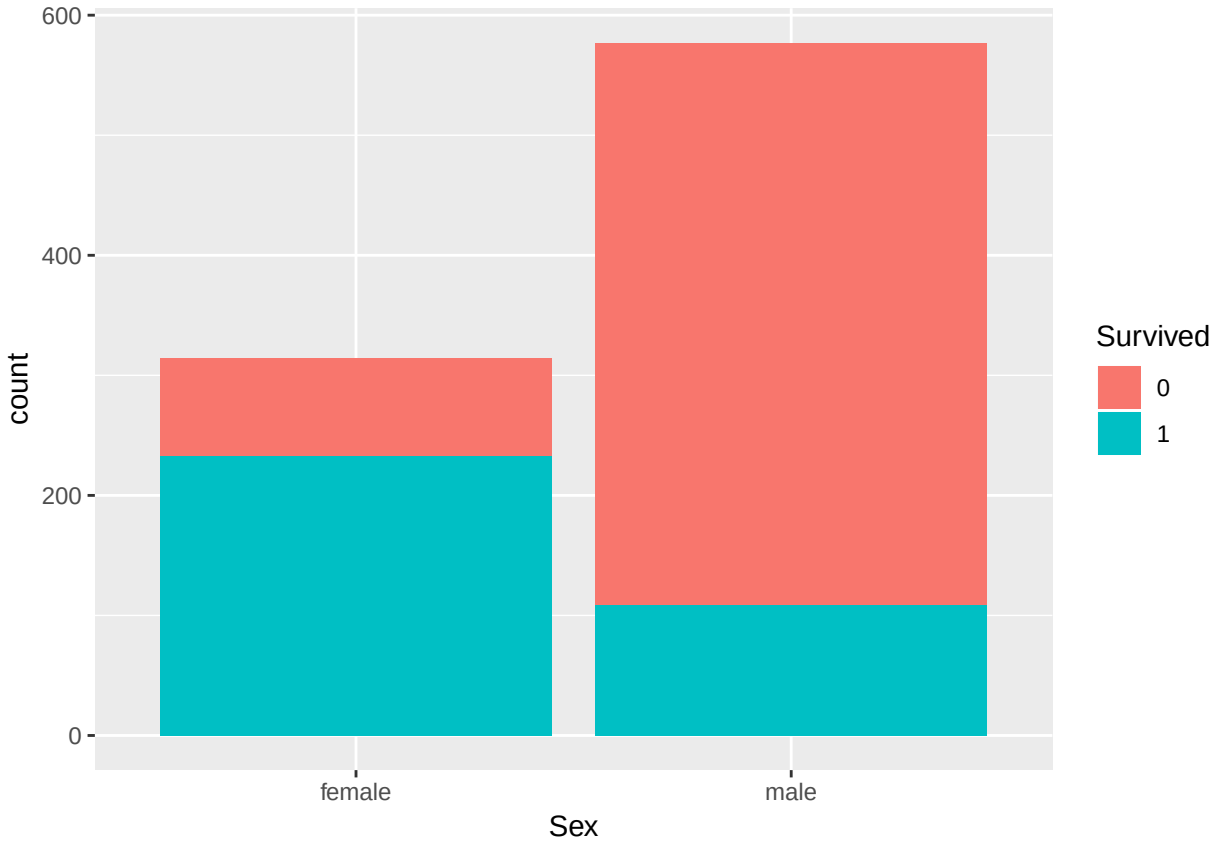


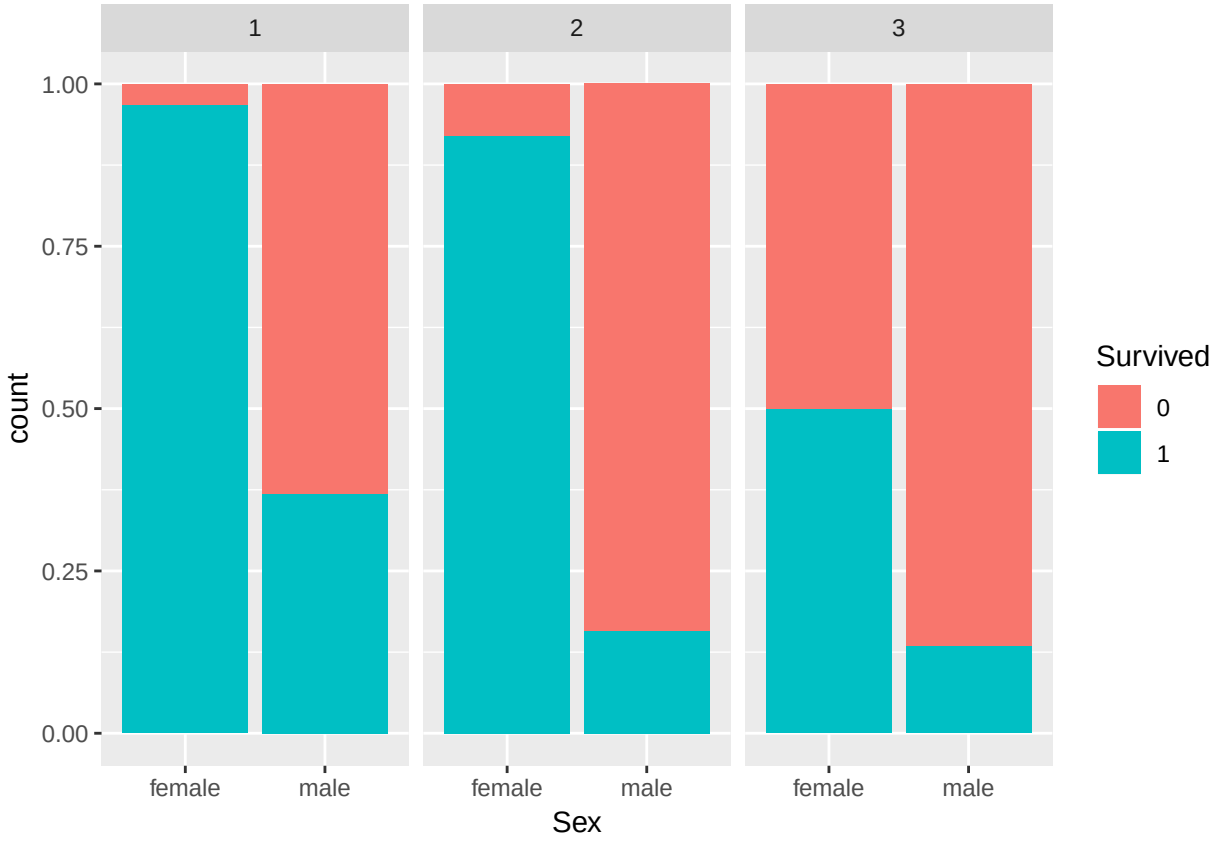
Tabla con los valores numéricos. La mayoría de mujeres sobrevivieron (233) mientras que en el caso contrario la mayoría de hombres perecieron (468).

```
table(titanic$Sex,titanic$Survived)
```

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

Gráfica comparativa en función del sexo y la clase en la que viajaban.

```
ggplot(data = titanic,aes(x=Sex,fill=Survived))+geom_bar(position="fill")+facet_wrap(~Pclass)
```

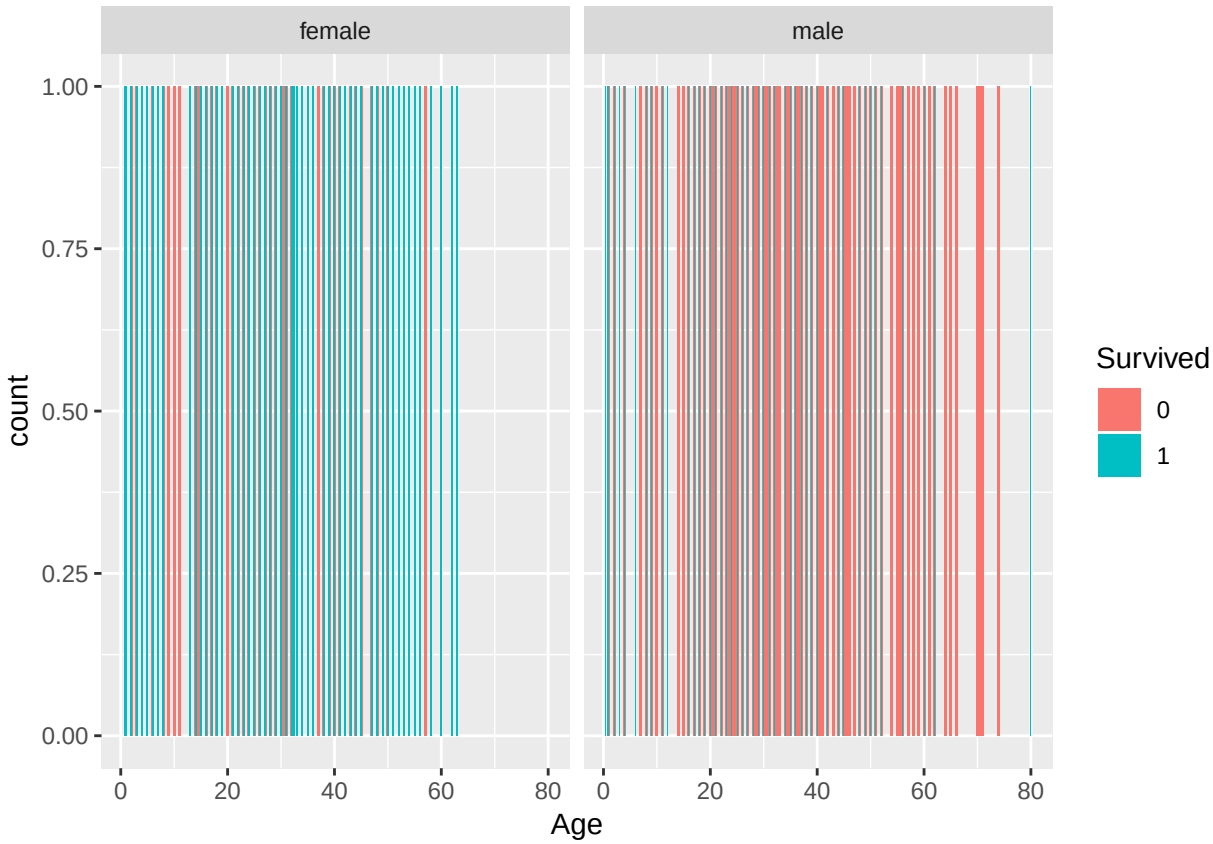


Reprentaciones en función de la edad de los viajeros.

```
ggplot(data = titanic,aes(x=Age,fill=Survived))+geom_bar(position="fill")+facet_wrap(~Sex)
```

```
## Warning: position_stack requires non-overlapping x intervals
```

```
## Warning: position_stack requires non-overlapping x intervals
```




```
#Creamos un subset del dataset original para analizar el caso de los niños menores de 10 años.
titanic2 <- subset(titanic, titanic$Age < 10)
#Y representamos su tabla.
table(titanic2$Age, titanic2$Survived)
```

```
##
##        0 1
##   0.42 0 1
##   0.67 0 1
##   0.75 0 2
##   0.83 0 2
##   0.92 0 1
##   1    2 5
##   2    7 3
##   3    1 5
##   4    3 7
##   5    0 4
##   6    1 2
##   7    2 1
##   8    2 2
##   9    6 2
```

Comparamos gráfico de frecuencia Survived - Parch

```
ggplot(data = titanic,aes(x=Parch,fill=Survived))+geom_bar()
```

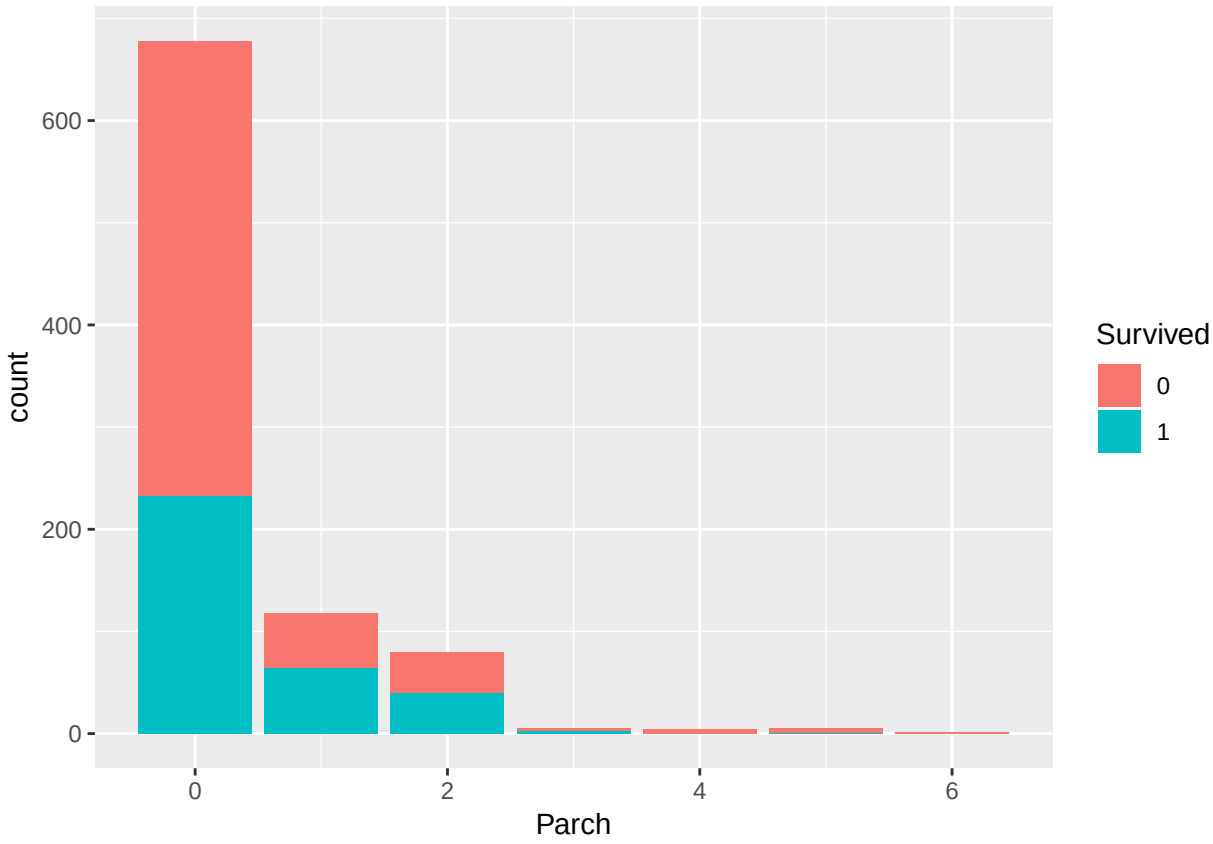


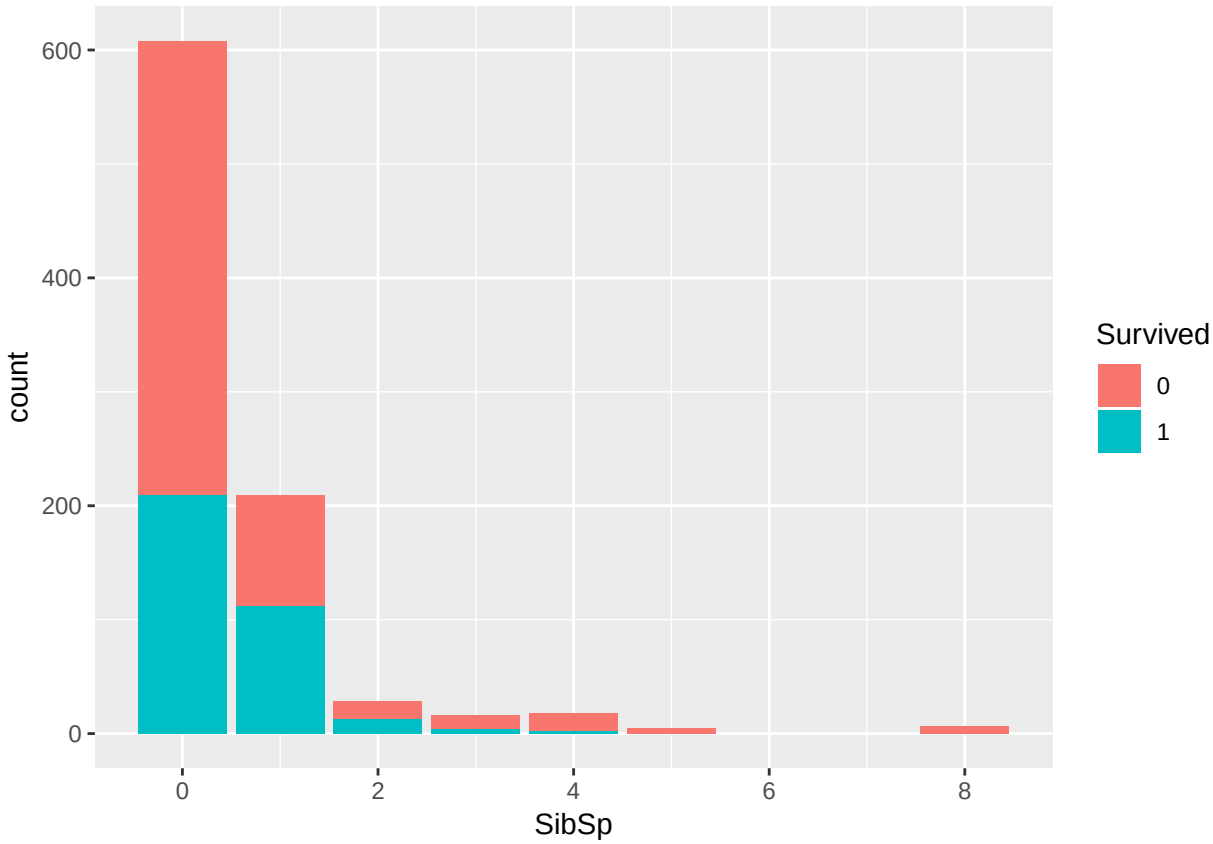
Tabla de frecuencia que muestra la comparativa Parch con Survived. Es decir si tienes 4 hijos o más las posibilidades de sobrevivir disminuyen drásticamente.

```
table(titanic$Parch,titanic$Survived)
```

```
##
##       0   1
##   0 445 233
##   1  53  65
##   2  40  40
##   3   2   3
##   4   4   0
##   5   4   1
##   6   1   0
```

Comparamos gráfico de frecuencia Survived - SibSp

```
ggplot(data = titanic,aes(x=SibSp,fill=Survived))+geom_bar()
```



Definimos una nueva variable llamada tamaño de familia, en la que sumamos la variable hermanos/cónyuge con hijos/padres.

```
# Construimos un atributo nuevo: family size.
titanic$FamSize <- titanic$SibSp + titanic$Parch +1;
titanic1 <- titanic
ggplot(data = titanic1[!is.na(titanic$FamSize),],aes(x=FamSize,fill=Survived))+geom_histogram(binwidth =
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```

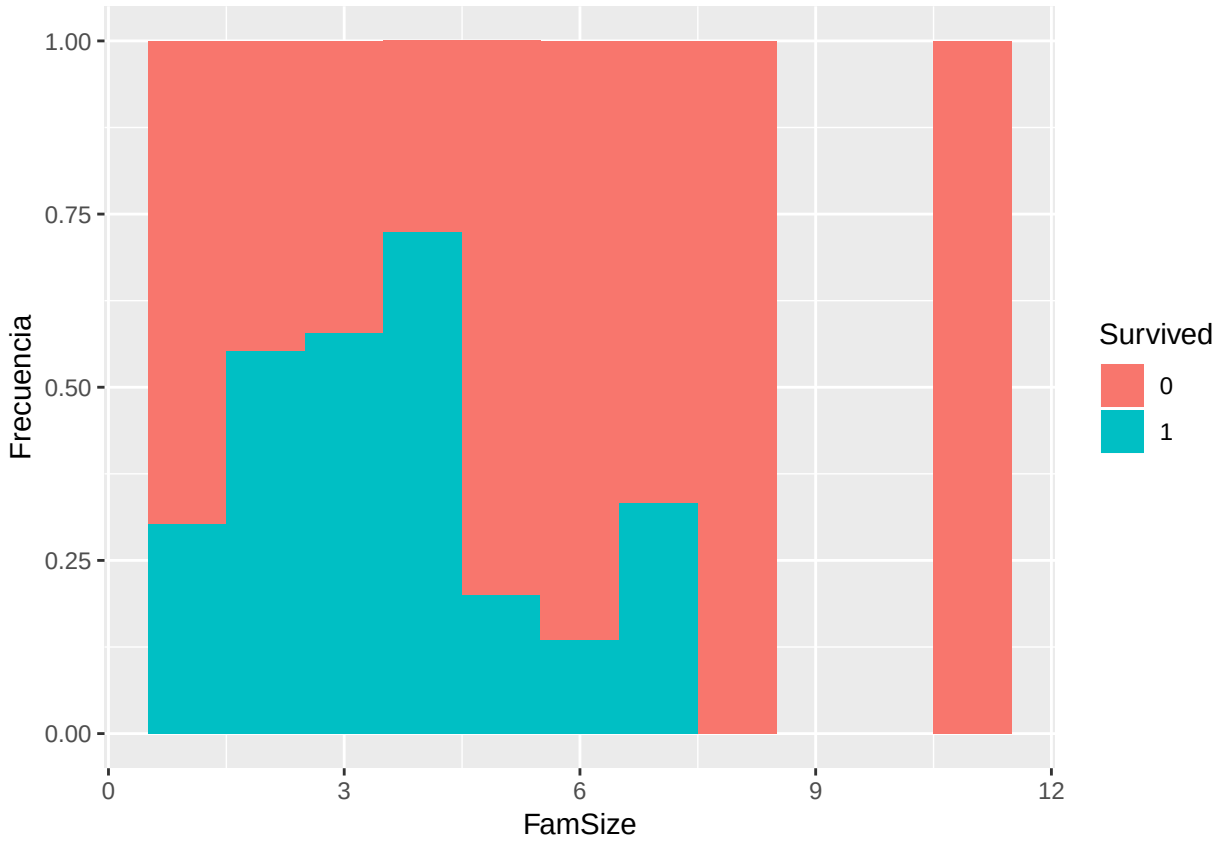


Tabla que muestra la relación entre el nuevo atributo y la superviviencia.

```
table(titanic$FamSize, titanic$Survived)
```

```
##
##        0   1
##   1  374 163
##   2   72  89
##   3   43  59
##   4    8  21
##   5   12   3
##   6   19   3
##   7    8   4
##   8    6   0
##   11   7   0
```

```
write.csv(titanic, "/Users/ramon/pra2_tipologia/titanic_data_clean.csv")
```

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Como podemos apreciar en los resultados obtenidos, había más hombres que mujeres en el barco, pero las posibilidades de sobrevivir siendo mujer eran bastante superiores que si eras hombre.

Por otro lado, las mujeres que viajaban en primera o segunda clase tenían bastantes posibilidades de supervivencia (+90%), sin embargo esta cifra se reducía a la mitad si viaja en tercera clase.

En el caso de los niños, los bebes de menos de un año corrieron mejor suerte que otros mejores de 10 años, ya que sí hubo caso en los que perdieron la vida (no es el caso en los bebés mencionados).

En general, vemos que si se viaja con entre 2 y 4 miembros de la familia hay más posibilidades de sobrevivir

que si se hace en solitario o con 5 o más miembros.

Podemos decir que sí responde al problema planteado, ya que a partir de estos resultados deducimos que si eres mujer tienes una mayor problabilidad de sobrevivir que si eres hombre, y que las clases 1 y 2 tienen bastantes más opciones que la 3. También hemos visto como si la familia que viaja se compone de entre 2 y 4 miembros las opciones de sobrevivir también son superiores que en otros casos, esto podría deberse a que en familias pequeñas los miembros se puede controlar y ayudar más fácilmente los unos a los otros.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código se adjunta en el fichero Rmd.