Práctica 2: Limpieza y análisis de datos

Juan Herranz Martin

7 de enero de 2020

Contents

1 Descripción del dataset

Este conocido dataset ha sido obtenido a través de la web *Kaggle* (https://www.kaggle.com/c/titanic/data), donde hemos seleccionado únicamente el conjunto de entrenamiento (train.csv) para la realización de la práctica, ya que este conjunto es lo suficientemente grande y posee la variable objetivo "Survived". El dataset contiene las siguientes variables:

- **PassengerId**: ID del pasajero.
- **Survived**: Si sobrevivió o no. 0=No, 1=Si.
- **Pclass**: El tipo de clase del ticket del pasajero. 1=Primera clase, 2=Segunda clase, 3=Tercera clase.
- **Name**: Nombre del pasajero.
- **Sex**: Sexo del pasajero.
- **Age**: Edad del pasajero.
- **SibSp**: Número de hermanos/cónyuges a bordo.
- **Parch**: Número de padres/hijos a bordo.
- **Ticket**: Número del ticket.
- **Fare**: Precio del ticket.
- **Cabin**: Cabina asignada al pasajero.
- **Embarked**: Puerto donde embarcó el pasajero. C=Cherbourg, Q=Queenstown, S=Southampton.

Este conjunto de datos es importante para poder analizar cuáles son las variables que más influyeron en la supervivencia de los pasajeros tras la catástrofe. Además, este conjunto contiene diferentes tipos de variables (tanto cuantitativas como cualitativas, numéricas y categóricas) que permitirán llevar a cabo varias tareas y métodos de análisis de datos diferentes, así como pruebas estadísticas.

Por tanto, se pretende responder a preguntas como: ¿Hay relación entre algunas de las variables?, ¿En qué medida influyen variables como la clase, el género o la edad del pasajero en la supervivencia?, ¿Podemos hacer predicciones sobre la supervivencia de ciertos pasajeros?, etc...

2 Integración y selección de datos

En primero lugar, cargamos los datos haciendo la lectura del fichero csv.

```
titanic<-read.csv("train.csv", header = T)
```

A continuación, vamos a prescindir de aquellas variables que menos información pueden aportar a los análisis que nos ocupan. Estas variables son: PassengerID, Name, Ticket, Fare, Cabin; que no van a ser usadas para responder a la ninguna de las preguntas planteadas.

```
titanic$PassengerId<-NULL
titanic$Name<-NULL
titanic$Ticket<-NULL
titanic$Fare<-NULL
titanic$Cabin<-NULL

#Vemos la estructura del conjunto
str(titanic)
```

```
## 'data.frame':    891 obs. of  7 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ 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 ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
#Tenemos 891 observaciones y 7 variables.
```

En este punto nos planteamos que una transformación lógica sería obtener el tamaño total de la familia del pasajero, a partir de la suma de las variables "SibSp", "Parch" mas el propio pasajero. Haremos uso de esta nueva variable para análisis posteriores.

```
#Llamamos Fam a esa nueva variable
titanic$Fam<-titanic$SibSp + titanic$Parch + 1

#prescindimos de las variables "SibSp" y "Parch"
titanic$SibSp<-NULL
titanic$Parch<-NULL

#hacemos un summary() para obtener un resumen algo más detallado
summary(titanic)
```

```
##     Survived          Pclass          Sex           Age        Embarked
##  Min.   :0.0000   Min.   :1.000   female:314   Min.   : 0.42    :  2
##  1st Qu.:0.0000   1st Qu.:2.000   male  :577   1st Qu.:20.12   C:168
##  Median :0.0000   Median :3.000                Median :28.00   Q: 77
##  Mean   :0.3838   Mean   :2.309                Mean   :29.70   S:644
##  3rd Qu.:1.0000   3rd Qu.:3.000                3rd Qu.:38.00
##  Max.   :1.0000   Max.   :3.000                Max.   :80.00
##                                                NA's   :177
##       Fam
##  Min.   : 1.000
##  1st Qu.: 1.000
##  Median : 1.000
##  Mean   : 1.905
##  3rd Qu.: 2.000
##  Max.   :11.000
##
```

Con esta información previa, prodecemos a llevar a cabo las tareas de limpieza

3 Limpieza de los datos

3.1 Ceros y/o nulos y/o elementos vacios

Comprobamos en qué variables se tienen valores cero, nulos o vacios:

```
#Valores cero
colSums(titanic==0)
```

```
## Survived   Pclass      Sex      Age Embarked      Fam
##      549        0        0       NA        0        0
```

```
#valores nulos
colSums(is.na(titanic))
```

```
## Survived   Pclass      Sex      Age Embarked      Fam
##        0        0        0      177        0        0
```

```
#valores vacios
colSums(titanic=="")
```

```
## Survived   Pclass      Sex      Age Embarked      Fam
##        0        0        0       NA        2        0
```

En el caso de los ceros, éstos no indican valores perdidos y/o desconocidos, si no que son todos valores posibles en las variables dadas.

En el caso de los valores desconocidos (NA) se tienen 177 de estos valores en la variable "Age". Ya que se trata de una variable numérica (cuya media y mediana no difieren demasiado) una solución podría ser imputar esos valores desconocidos con la media de edad total según el sexo.

En el caso de valores vacios, tenemos 2 de estos valores en la variable Embarked. Una solución sería sustituir estos dos valores (ya que son pocos) por la categoría más frecuente, que es, con diferencia, la categoría "S".

Llevamos a cabo las soluciones que acabamos de describir:

```
titanic$Age[is.na(titanic$Age) & titanic$Sex=="male"] <-
  round(mean(titanic$Age[titanic$Sex=="male"],na.rm = T))
titanic$Age[is.na(titanic$Age) & titanic$Sex=="female"] <-
  round(mean(titanic$Age[titanic$Sex=="female"],na.rm = T))

#convertimos "Embarked" a caracter para imputar y reconvertimos a factor para eliminar el factor ""
titanic$Embarked<-as.character(titanic$Embarked)
titanic$Embarked[titanic$Embarked==""]="S"
titanic$Embarked<-as.factor(titanic$Embarked)

#veamos como ha quedado la estructura y si tiene sentido cambiar el formato de alguna variable
str(titanic)
```

```
## 'data.frame':    891 obs. of  6 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 31 54 2 27 14 ...
##  $ Embarked: Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ Fam     : num  2 2 1 2 1 1 1 5 3 2 ...
```

```
#conviene factorizar las variables "Survived" y "Pclass"
titanic$Survived<-as.factor(titanic$Survived)
titanic$Pclass<-as.factor(titanic$Pclass)

str(titanic)
```

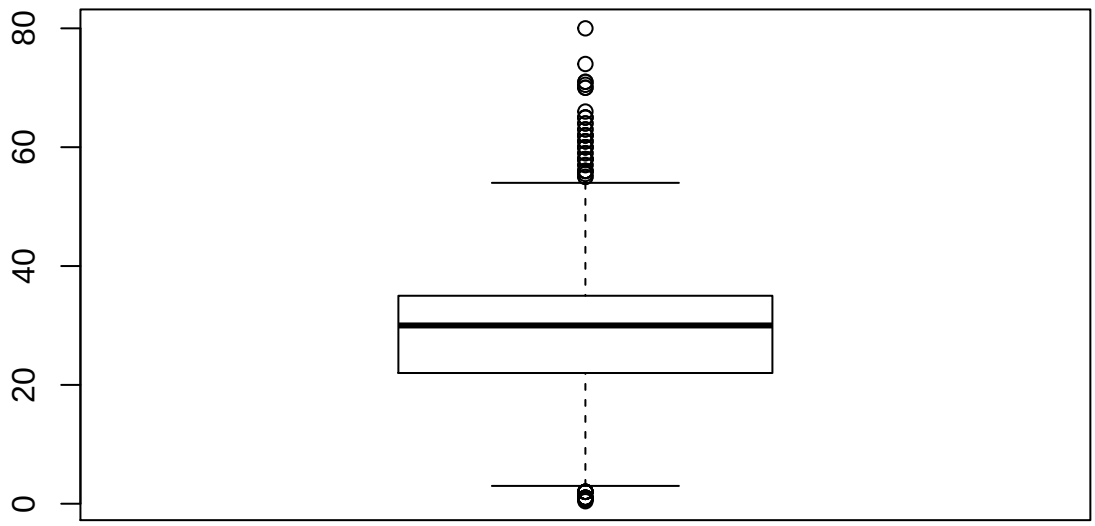
```
## 'data.frame':    891 obs. of  6 variables:
##  $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 31 54 2 27 14 ...
##  $ Embarked: Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ Fam     : num  2 2 1 2 1 1 1 5 3 2 ...
```

Todo parece indicar que ya disponemos de las variables en los formatos adecuados y sin valores desconocidos o nulos. Vamos ahora a identificar y tratar los posibles outliers.

3.2 Identificación y análisis de valores extremos.

Analizamos los valores extremos de las variables numéricas "Age" y "Fam". Para ello hacemos uso de diagramas de cajas y de la función boxplot.stats().

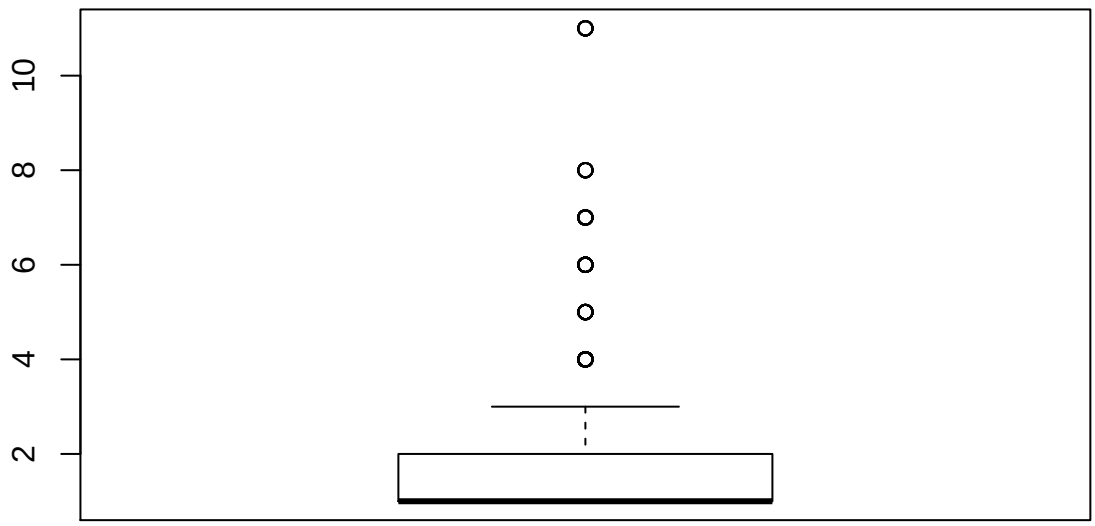
```
#Outliers en "Age"
boxplot(titanic$Age)
```




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

```
#Outliers en "Fam"
boxplot(titanic$Fam)
```




```
boxplot.stats(titanic$Fam)$out
```

```
##  [1]  5  7  6  5  7  6  4  6  4  8  6  7  8  4  5  6  4  7  5 11  6  6  6
## [24]  5 11  7  4 11  5  7  7  6  6  4  4  5 11  6  6  5  8  4  5  4  5  6
## [47]  6  4  4  4  4  8  5  4  4  7  7  5  4  4  7  4  4  6  6  6  4  8  8
## [70]  4  6  4  5  5  4  4  5  4  6  4 11  4  7  6  6 11  7  4 11  6  4
```

Para analizar mejor estos valores vamos a obtener los valores máximo y mínimo de estas variables y vamos a interpretar si son valores lógicos o posibles dentro de un rango realista.

```
max(titanic$Age)
```

```
## [1] 80
```

```
min(titanic$Age)
```

```
## [1] 0.42
```

```
max(titanic$Fam)
```

```
## [1] 11
```

```
min(titanic$Fam)
```

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

Los valores máximo y mínimo de la variable familia son valores lógicos y posibles dentro del rango (familias entre 1 y 11 miembros), por lo que no vamos a tratar los outliers de esta variable.

El valor máximo obtenido para la edad también es un valor lógico y posible (80 años). Sin embargo, el valor mínimo obtenido (0.42) choca un poco con el resto de valores aunque probablemente se trate del valor en

meses proporcional al año. Para no tener este tipo de valores y conseguir que la edad sea una variable en el rango [1,80], vamos redondear por arriba todos los valores inferiores a uno, de forma que todos los bebés de varios meses queden registrados como valor 1 en la edad, siendo éste el valor mínimo para esta variable:

```
titanic$Age[titanic$Age<1]<-1

#Hacemos un summary para verificar los resultados
summary(titanic)
```

```
##  Survived Pclass      Sex           Age        Embarked      Fam
##  0:549    1:216   female:314   Min.   : 1.00   C:168    Min.   : 1.000
##  1:342    2:184   male  :577   1st Qu.:22.00   Q: 77    1st Qu.: 1.000
##           3:491                Median :30.00   S:646    Median : 1.000
##                                Mean   :29.78            Mean   : 1.905
##                                3rd Qu.:35.00            3rd Qu.: 2.000
##                                Max.   :80.00            Max.   :11.000
```

En este punto podemos dar por finalizado el proceso de limpieza de los datos.

4 Análisis de los datos

4.1 Selección de los grupos de datos a analizar/comparar

Seleccionamos ahora los grupos incluidos en el conjunto que podremos analizar o comparar.

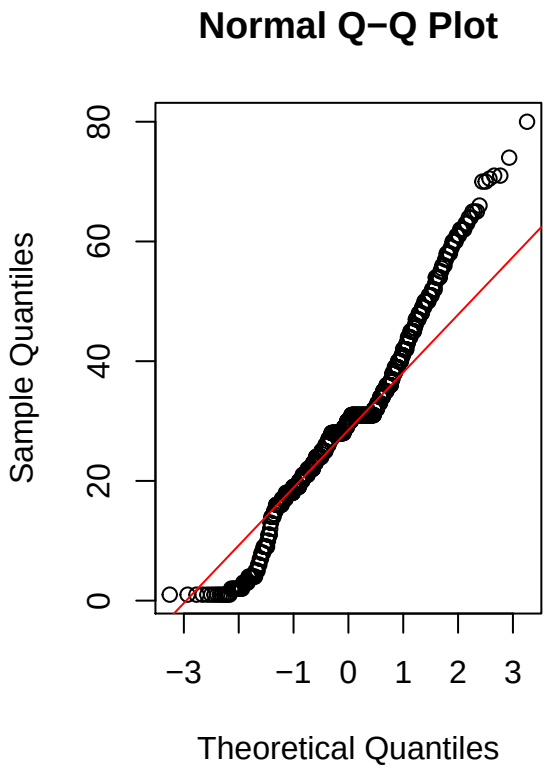
```
#Grupos según la clase
titanic.primera<-titanic[titanic$Pclass==1,]
titanic.segunda<-titanic[titanic$Pclass==2,]
titanic.tercera<-titanic[titanic$Pclass==3,]

#Grupos según el sexo
titanic.hombre<-titanic[titanic$Sex=="male",]
titanic.mujer<-titanic[titanic$Sex=="female",]

#Grupos según la edad
#En este caso discretizamos esta variable en 3 categorías:
titanic.menores<-titanic[titanic$Age<18,]
titanic.adultos<-titanic[titanic$Age>=18 & titanic$Age<65,]
titanic.mayores<-titanic[titanic$Age>=65,]

#Grupos según la familia
#En este caso discretizamos esta variable en 4 categorías:
titanic.individual<-titanic[titanic$Fam==1,]
titanic.pareja<-titanic[titanic$Fam==2,]
titanic.familia<-titanic[titanic$Fam>2 & titanic$Fam<5,]
titanic.numerosa<-titanic[titanic$Fam>=5,]
```

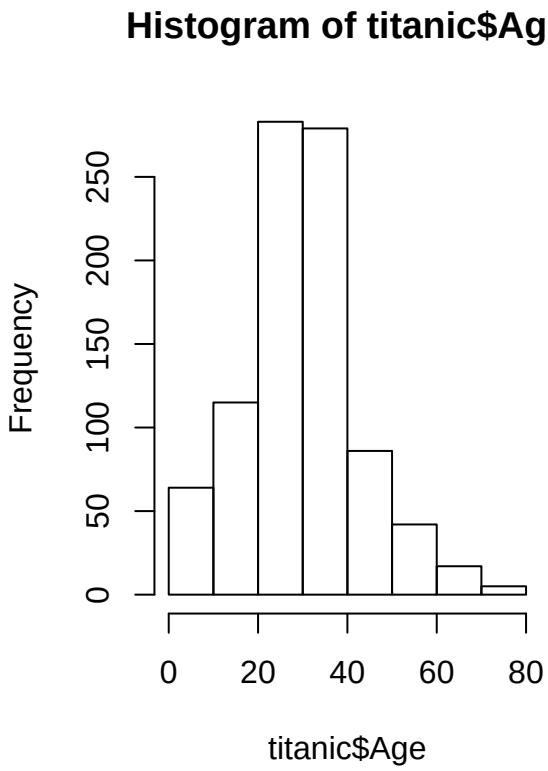
4.2 Comprobación de la normalidad y homogeneidad de la varianza

Vamos ahora a comprobar si las variables cuantitativas ("Age") provienen de una población que sigue una distribución normal. Para ello, veremos con ayuda de la visualización del histograma y de las curvas Q-Q si la variable puede ser normalizada y llevando a cabo el test de normalidad de Shapiro-Wilk comprobaremos si realmente proviene de una distribución normal.

```
par(mfrow=c(1,2))
qqnorm(titanic$Age)
```

```
qqline(titanic$Age,col="red")
hist(titanic$Age)
```





Vemos que la variable edad puede ser normalizada. Además, se puede aplicar el teorema del límite central ya que el número de muestras es suficientemente grande (n>30), haciendo un cambio de variable sobre la media cuya distribución se aproximará a una distribución normal estándar.

Comprobamos ahora la normalidad con el test de Shapiro-Wilk

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

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Age
## W = 0.96407, p-value = 5.333e-14
```

La variable "Age" no está normalizada ya que el p-valor obtenido ($5.33 \cdot 10^{-14}$) es inferior al nivel de significancia($\alpha = 0.05$), por lo que se rechaza la hipótesis nula de normalidad de la población.

Comprobamos ahora la homogeneidad de la varianza en esa misma variable. Ya que no hemos asegurado que la distribución de la población siga una distribución normal, debemos hacer uso de algún test de homogeneidad de la varianza que no sea sensible a la falta de normalidad, como puede ser el test de Fligner-Killeen, que hace uso de la mediana. Comprobaremos la homogeneidad en cuanto al sexo de los pasajeros, comparando así ambas muestras:

```
fligner.test(Age ~ Sex, data=titanic)
```

```
##
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Sex
## Fligner-Killeen:med chi-squared = 0.33349, df = 1, p-value =
## 0.5636
```

Se obtiene un p-valor (0.5636) superior al nivel de significancia, por lo que aceptamos la hipótesis nula de varianzas homogeneas a partir de esas dos muestras.

4.3 Aplicación de pruebas estadísticas

4.3.1 Contraste de hipótesis sobre la independencia de variables categóricas

La primera prueba estadística será la aplicación del contraste de hipótesis de la independencia de algunas variables categóricas del conjunto ("Sex" y "Pclass") para determinar si éstas son o no estadísticamente independientes. Para ello, vamos a considerar que la hipótesis nula (H_0) es que las variables son independientes frente a la hipótesis alternativa (H_1) de que no lo son. Para contrastarlo, hacemos uso de pruebas no paramétricas como el test χ^2 de Pearson y el test exacto de Fisher.

```
#creamos las tablas de contingencia
t1<-table(titanic$Survived,titanic$Sex)
t1
```

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

```
t2<-table(titanic$Survived,titanic$Pclass)
t2
```

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

```
#aplicamos los tests sobre las tablas
chisq.test(t1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  t1
## X-squared = 260.72, df = 1, p-value < 2.2e-16
```

```
chisq.test(t2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  t2
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

```
fisher.test(t1)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t1
```

```
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.0575310 0.1138011
## sample estimates:
## odds ratio
## 0.08128333
```

```
fisher.test(t2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  t2
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Los p-valor obtenidos de ambos test (inferiores al nivel de significancia $\alpha = 0.05$), así como el estadístico χ^2 indican que se tiene que rechazar la hipótesis nula en ambos casos. Por tanto, según las pruebas estadísticas, las variables "Sex" y "Pclass" tienen influencia sobre la variable "Survived".

4.3.2 Regresión logística.

En este caso vamos a llevar a cabo una regresión logística (ya que la variable respuesta "Survived" es dicotómica) seleccionando la variable "Age" para construir un primer modelo. Haremos uso de la función de modelos lineales generalizados glm() especificando que la familia de la distribución es la binomial.

```
log_model <- glm(data=titanic, Survived ~ Age, family = binomial)
summary(log_model)
```

```
##
## Call:
## glm(formula = Survived ~ Age, family = binomial, data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1355  -0.9919  -0.9363   1.3576   1.6816
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.086246   0.172617  -0.500   0.6173
## Age         -0.013113   0.005409  -2.424   0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance: 1180.7  on 889  degrees of freedom
## AIC: 1184.7
##
## Number of Fisher Scoring iterations: 4
```

El p-valor obtenido para la variable "Age" ha sido 0.0153, lo que significa que esta variable es moderadamente significativa para el modelo, ya que el p-valor se encuentra entre los valores 0.01 y 0.05.

Vamos a binarizar la variable edad y familia para construir el modelo, y a añadir también las variables clase

y sexo al modelo para mejorarlo e interpretar los resultados.

```
titanic$Age_bin[titanic$Age<18]<-0
titanic$Age_bin[titanic$Age>=18]<-1

titanic$Fam_bin[titanic$Fam<5]<-0
titanic$Fam_bin[titanic$Fam>=5]<-1

log_model2 <- glm(data=titanic, Survived ~ Age_bin + Pclass + Sex + Fam_bin, family = binomial)
summary(log_model2)
```

```
##
## Call:
## glm(formula = Survived ~ Age_bin + Pclass + Sex + Fam_bin, family = binomial,
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8394  -0.6614  -0.4263   0.6514   2.5766
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.0132     0.3899  10.293  < 2e-16 ***
## Age_bin      -1.5348     0.2965  -5.176 2.27e-07 ***
## Pclass2      -1.0359     0.2556  -4.052 5.07e-05 ***
## Pclass3      -1.9800     0.2256  -8.778  < 2e-16 ***
## Sexmale      -2.8509     0.1996 -14.283  < 2e-16 ***
## Fam_bin      -2.4647     0.4320  -5.705 1.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  771.51  on 885  degrees of freedom
## AIC: 783.51
##
## Number of Fisher Scoring iterations: 5
```

Vemos cómo ahora los p-valor de todas las variables son bastante inferiores a 0.05, por lo que todas las variables son muy significativas para el modelo. Podemos interpretar los parámetros obtenidos por el modelo de la siguiente forma:

- El coeficiente estimado para la variable edad es -1.5348, lo que significa que para un pasajero mayor de edad (>=18) disminuye el logaritmo del odds ratio ($\frac{p}{1-p}$) de la variable Survived un promedio de 1.5348 unidades. Tomando la exponencial de esta cantidad:

```
exp(1.5348)
```

```
## [1] 4.640397
```

se deduce que si el pasajero es mayor de edad, el odds ratio disminuye un promedio de 4.640397. Esto significa que los pasajeros mayores de edad tienen unas 4.64 veces más de riesgo de morir que los menores de edad.

- El coeficiente estimado para la variable de la segunda clase en relación a la primera es -1.0359, lo que significa que para un pasajero que viaja en segunda clase disminuye el logaritmo del odds ratio de la variable Survived un promedio de 1.0359 unidades. Tomando la exponencial de esta cantidad:

```
exp(1.0359)
```

```
## [1] 2.817641
```

se deduce que si el pasajero viaja en segunda clase, el odds ratio disminuye un promedio de 2.817641. Esto significa que los pasajeros de segunda clase tienen unas 2.817641 veces más de riesgo de morir que los de primera clase.

- El coeficiente estimado para la variable de la tercera clase en relación a la primera es -1.98, lo que significa que para un pasajero que viaja en tercera clase disminuye el logaritmo del odds ratio de la variable Survived un promedio de 1.98 unidades. Tomando la exponencial de esta cantidad:

```
exp(1.98)
```

```
## [1] 7.242743
```

se deduce que si el pasajero viaja en tercera clase, el odds ratio disminuye un promedio de 7.242743. Esto significa que los pasajeros de tercera clase tienen unas 7.242743 veces más de riesgo de morir que los de primera clase.

- El coeficiente estimado para la variable sexo es -2.8509, lo que significa que para un pasajero hombre disminuye el logaritmo del odds ratio de la variable Survived un promedio de 2.8509 unidades. Tomando la exponencial de esta cantidad:

```
exp(2.8509)
```

```
## [1] 17.30335
```

se deduce que si el pasajero es hombre, el odds ratio disminuye un promedio de 17.30335. Esto significa que los hombres tienen unas 17.30335 veces más de riesgo de morir que las mujeres.

- El coeficiente estimado para la variable familia es -2.4647, lo que significa que para un pasajero con familia numerosa disminuye el logaritmo del odds ratio de la variable Survived un promedio de 2.4647 unidades. Tomando la exponencial de esta cantidad:

```
exp(2.4647)
```

```
## [1] 11.75995
```

se deduce que si el pasajero tiene familia numerosa, el odds ratio disminuye un promedio de 11.75995. Esto significa que los pasajeros con familia numerosa tienen unas 11.75995 veces más de riesgo de morir que los pasajeros sin familia numerosa.

4.3.3 Predicción

Finalmente, podemos hacer uso del modelo anterior para predecir la probabilidad de sobrevivir de pasajeros con ciertas características. Vamos a suponer un primer pasajero que tiene 30 años (mayor de edad), viaja en segunda clase, es hombre y viaja con su mujer y cuatro hijos (tiene familia numerosa). Hacemos uso de la función predict() especificando el argumento type="response" para obtener directamente la probabilidad.

```
pred<-predict(log_model2, data.frame(Age_bin=1,Pclass="2",Sex="male",Fam_bin=1),type = "response")
pred
```

```
##          1
## 0.02036786
```

Este hombre tiene una probabilidad de sobrevivir del 2.04% según el modelo.

Vamos a suponer ahora otro pasajero: de 8 años (menor de edad), que viaja en primera clase, es mujer y viaja con sus dos padres (no tiene familia numerosa).

```
pred2<-predict(log_model2, data.frame(Age_bin=0,Pclass="1",Sex="female",Fam_bin=0),type = "response")
pred2
```

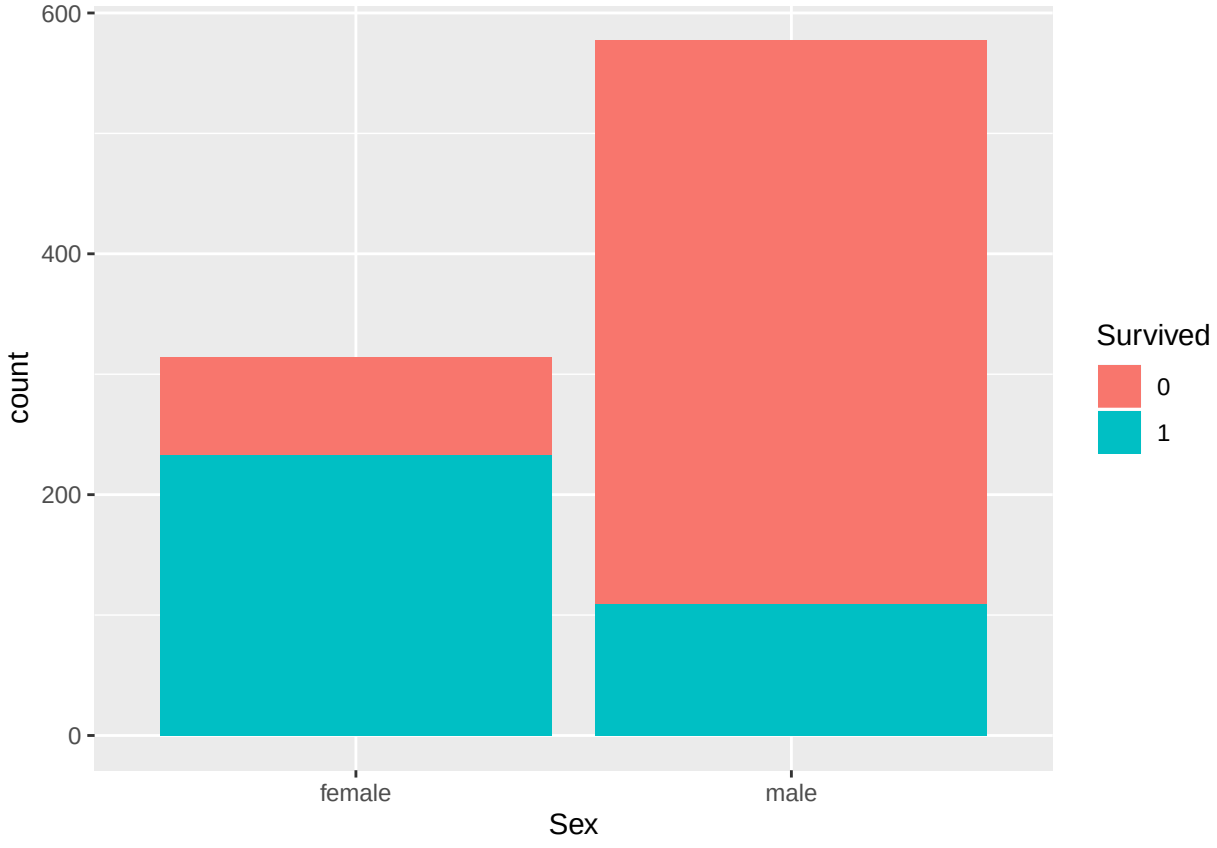
```
##         1
## 0.9822446
```

Esta niña tiene una probabilidad de sobrevivir del 98.22% según el modelo.

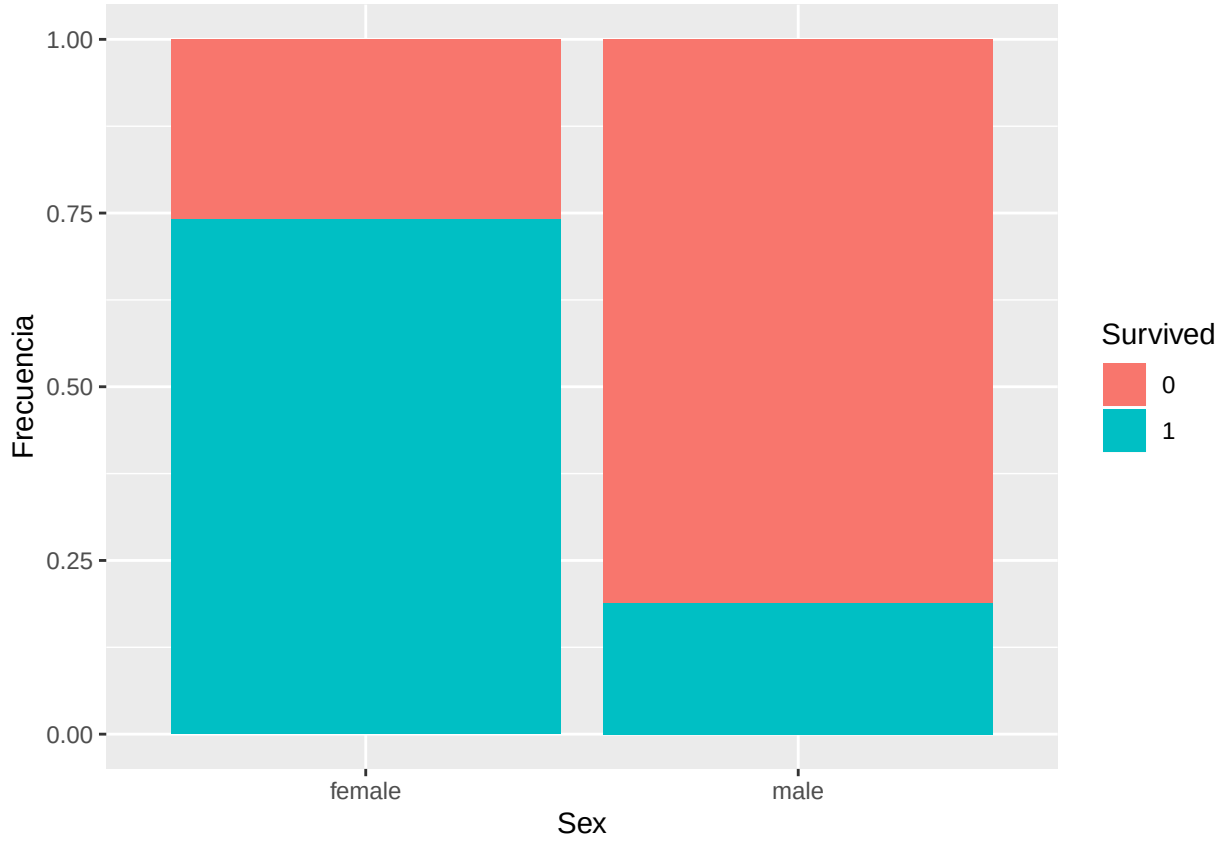
5 Representación gráfica de los resultados.

Vamos a visualizar las relaciones entre las diferentes variables del conjunto para poder ver algunos de los resultados anteriores. Para ello, haremos uso del paquete ggplot2. En concreto, vamos a representar las relaciones de la variable "Survived" con: - "Sex":

```
library(ggplot2)
rows=dim(titanic)[1]

#Representación del número total
ggplot(data=titanic[1:rows,],aes(x=Sex,fill=Survived))+geom_bar()
```

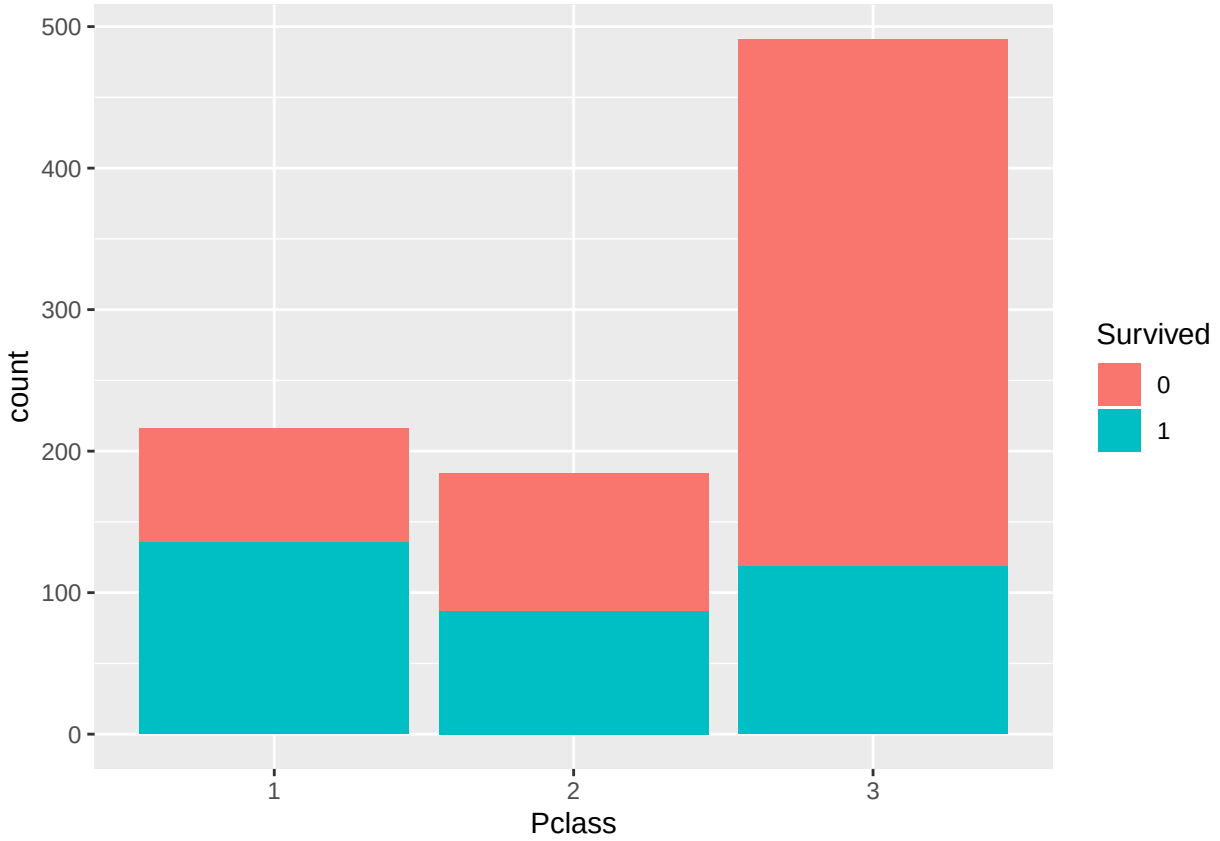



```
#Representación de la frecuencia proporcional
ggplot(data=titanic[1:rows,],aes(x=Sex,fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```

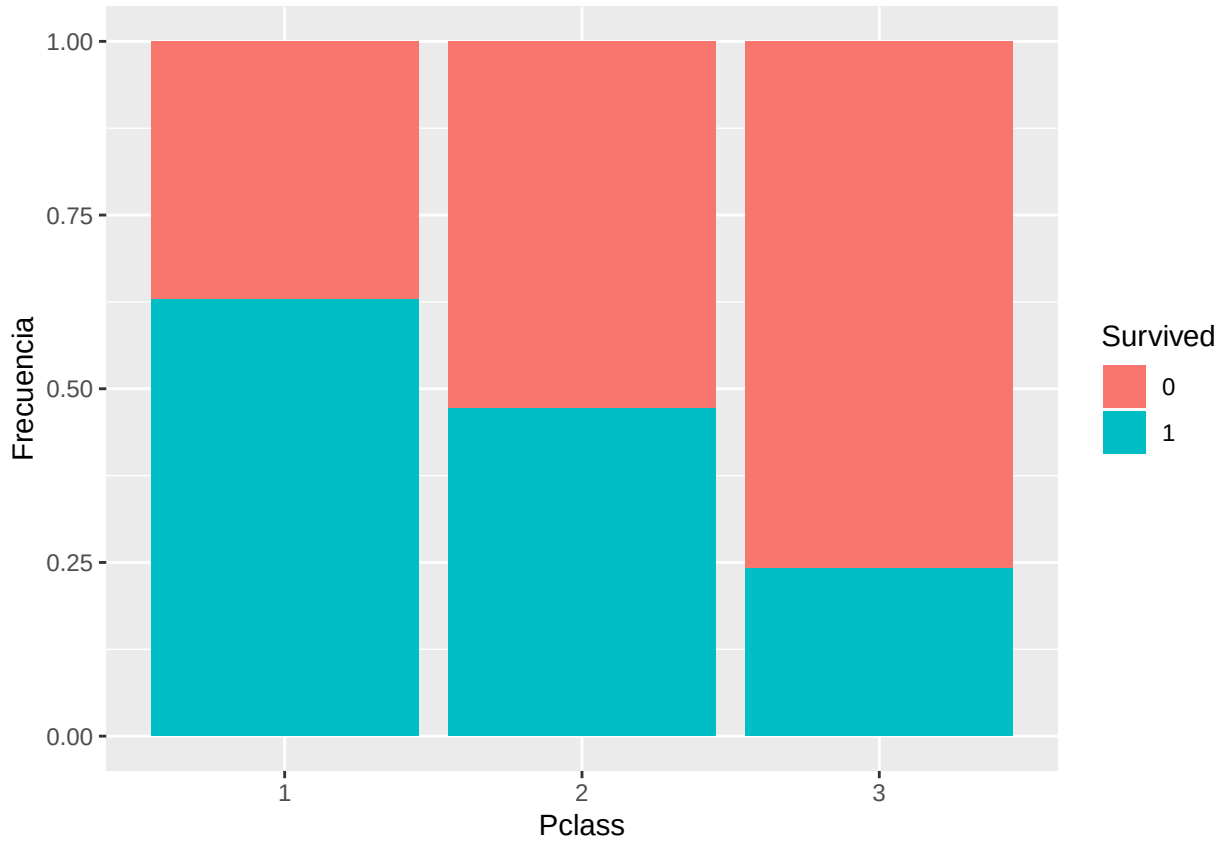


- “Pclass”

```
#Representación del número total
ggplot(data=titanic[1:rows,],aes(x=Pclass,fill=Survived))+geom_bar()
```

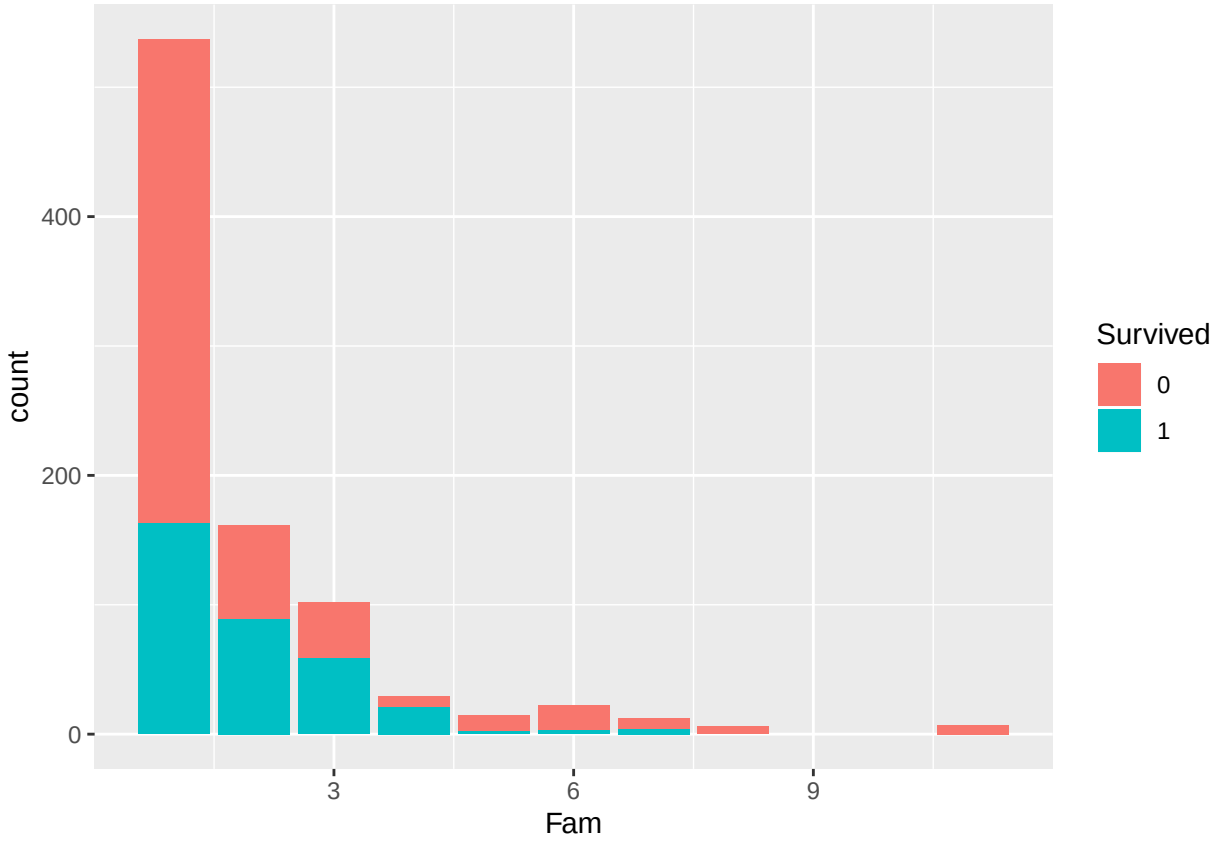



```
#Representación de la frecuencia proporcional
ggplot(data=titanic[1:rows,],aes(x=Pclass,fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```

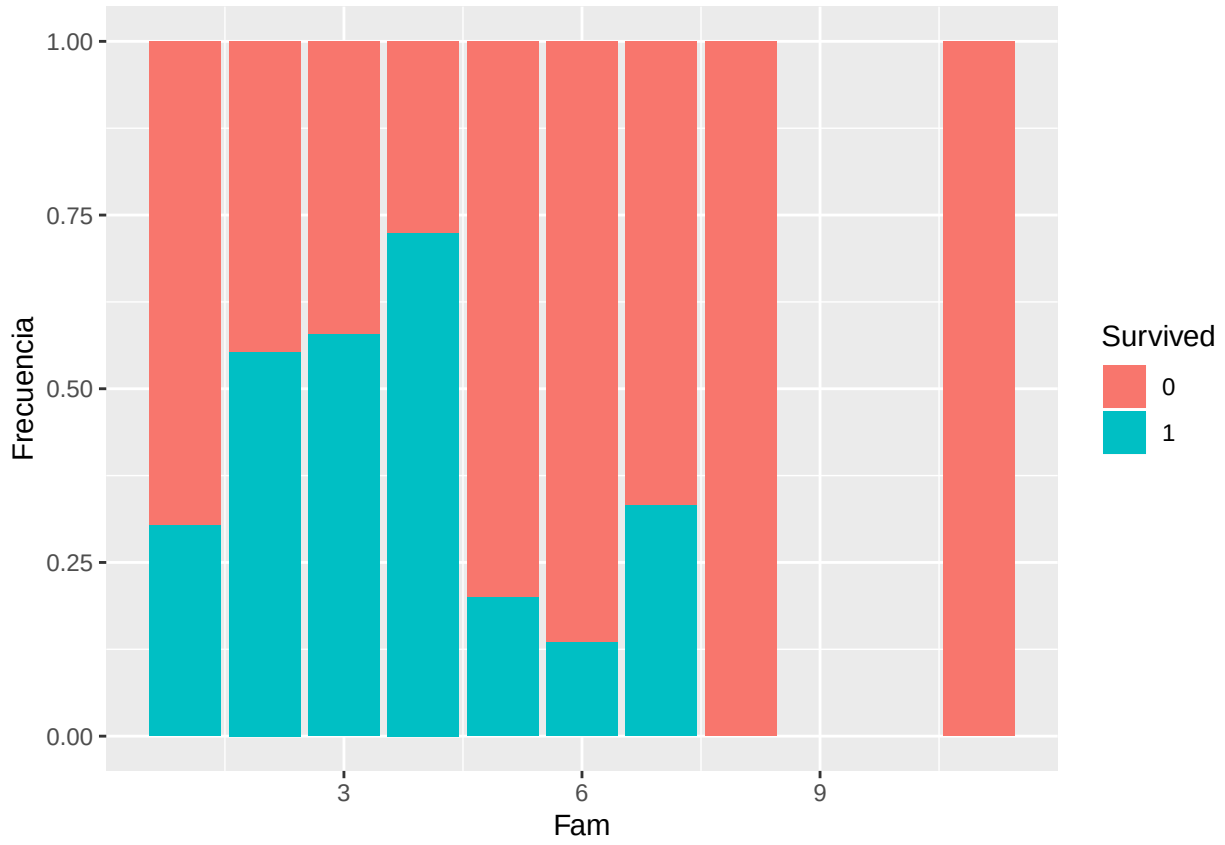


- "Fam"

```
#Representación del número total
ggplot(data=titanic[1:rows,],aes(x=Fam,fill=Survived))+geom_bar()
```

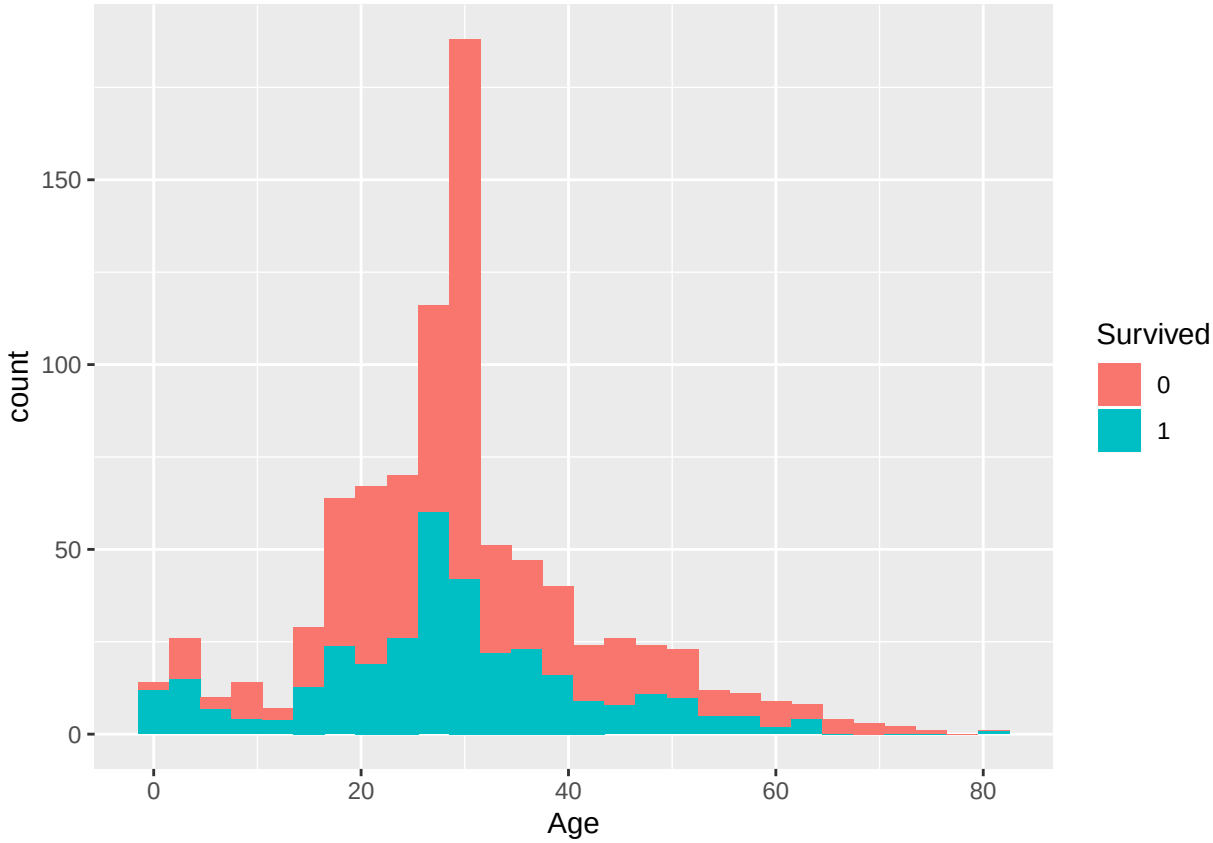



```
#Representación de la frecuencia proporcional
ggplot(data=titanic[1:rows,],aes(x=Fam,fill=Survived))+geom_bar(position="fill")+ylab("Frecuencia")
```



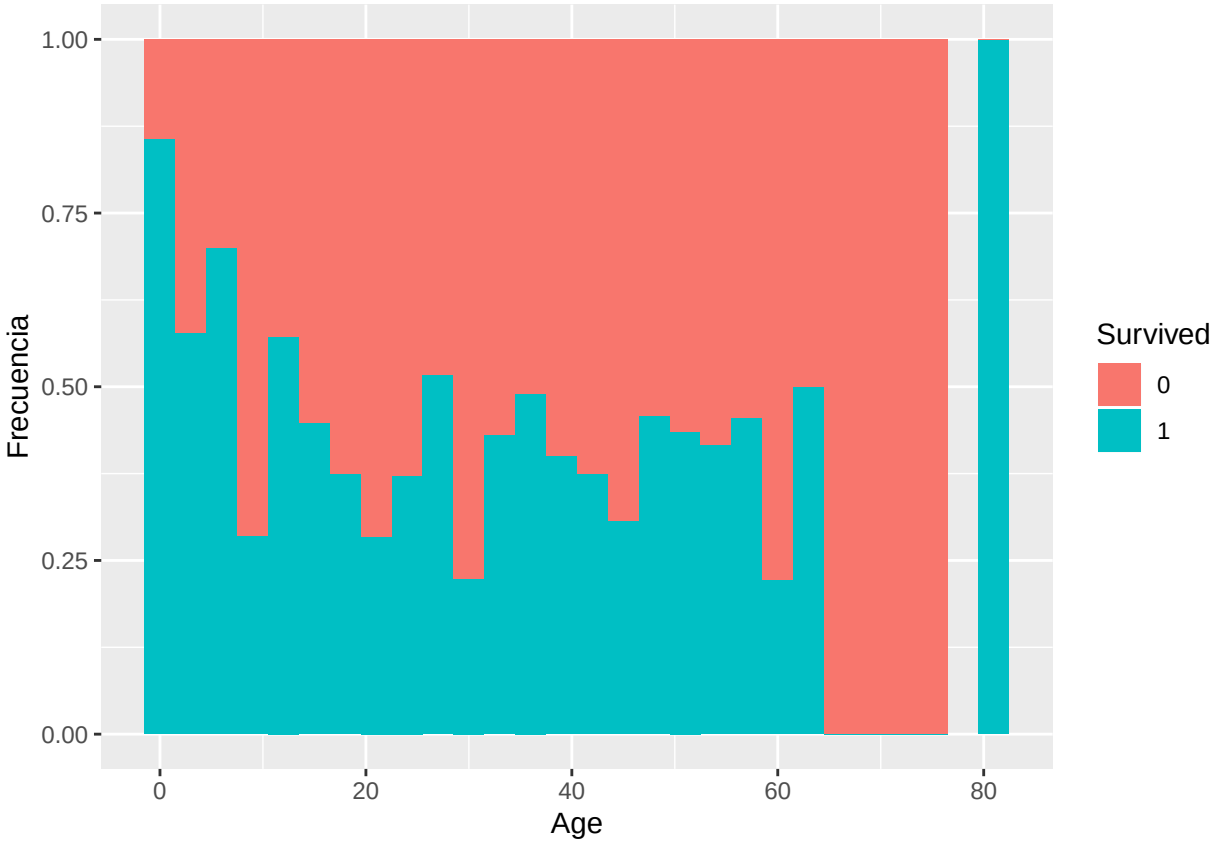
- "Age"

```
#Representación del número total
ggplot(data=titanic[1:rows,],aes(x=Age,fill=Survived))+geom_histogram(binwidth=3)
```




```
#Representación de la frecuencia proporcional
ggplot(data=titanic[1:rows,],aes(x=Age,fill=Survived))+geom_histogram(binwidth=3, position="fill")+ylab
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

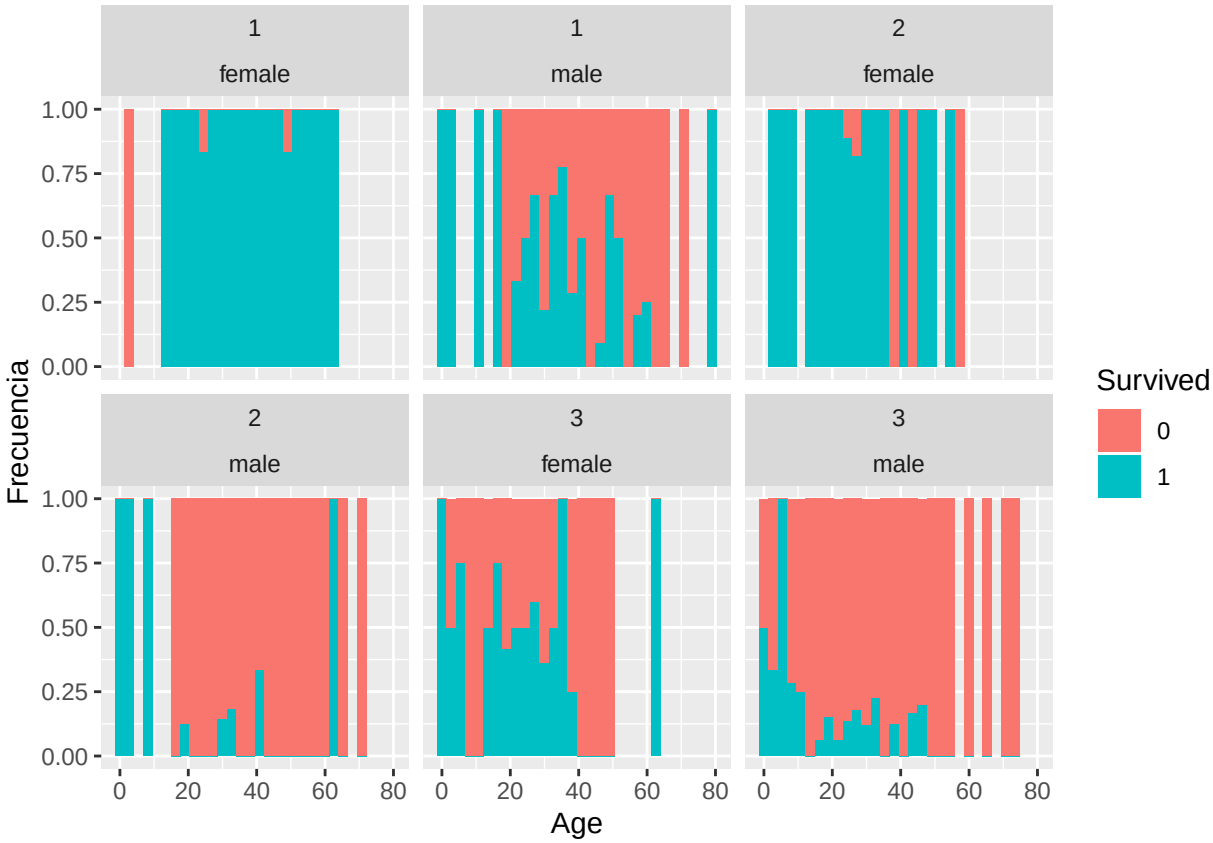


Tambien podemos representar más de dos variables a la vez, con la función facet_wrap. Por ejemplo, representamos "Survived" frente a "Age", "Pclass" y "Sex"

```
ggplot(data=titanic[1:rows,],aes(x=Age,fill=Survived))+geom_histogram(position = "fill")+
  ylab("Frecuencia")+facet_wrap(~Pclass+Sex)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 98 rows containing missing values (geom_bar).
```



6 Conclusiones

En vista de los resultados de las pruebas estadísticas, se pueden extraer las siguientes conclusiones:

- Las variables categóricas “Sex” y “Pclass” no son independientes de la variable “Survived”, según el contraste de hipótesis llevado a cabo.
- Según el modelo de regresión logística: los pasajeros hombres son los que menos probabilidades tienen de sobrevivir, seguidos de los pasajeros con familia numerosa, de los que viajan en tercera clase, de los mayores de edad y de los de segunda clase. El perfil de pasajero con mayor probabilidad de supervivencia sería: mujer, sin familia numerosa, de primera clase y menor de edad.

Con ayuda de las representaciones gráficas, se verifican los resultados del modelo logístico y se observan otros hechos:

- Se observa un mayor número de pasajeros hombres que mujeres y que la frecuencia relativa de mujeres supervivientes es mayor que la de los hombres.
- Se observa un mayor número de pasajeros de tercera clase y que la frecuencia relativa de supervivencia de éstos es la menor, seguidos de la segunda y la primera clase.
- Se observa un mayor número de pasajeros con un sólo individuo por familia, que la mayor frecuencia relativa de supervivencia la tienen los pasajeros con 4 miembros en la familia y que la menor la tienen los de 6 miembros. Tambien observamos que no hay casos de familias con 9 o 10 miembros. No sobrevivió ninguna familia con más de 7 miembros.
- Se observa un mayor número de pasajeros de unos 30 años y una distribución en la edad similar a una normal. La frecuencia relativa de supervivencia más alta es la de los pasajeros con más de 80 años, aunque fueron muy pocos. La siguiente frecuencia más alta es la de los bebés de un año o menos. No

sobrevivió ningún pasajero mayor de 65 y menor de 80 años. La siguiente frecuencia más baja la tienen los pasajeros de alrededor de 30 y 60 años. No sobrevivió ningún pasajero mayor de 65 y menor de 80 años.

- Observamos en la representación combinada, una gran proporción de supervivencia para las mujeres de primera clase y una gran proporción de no supervivencia para los hombres de tercera clase. Las mujeres y los niños vemos que tienen siempre más probabilidad de sobrevivir que el resto aunque varía según la clase.

Por tanto, la famosa órden del capitán del Titanic: "las mujeres y niños primero" parece ser cierta en este caso, pero matizable: "las mujeres y niños ricos, primero".

7 Código fuente.

El código fuente en R puede encontrarse en mi repositorio de *Github*: https://github.com/jherranzma/Practica_2