# Práctica 2: Limpieza y análisis de datos

Javier Fortea

5/29/2021



# Práctica 2: Limpieza y análisis de datos

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

El objetivo de esta actividad es el tratamiento de un dataset. En mi caso he escogido el siguiente dataset de Kaggle propuesto en la actividad:

- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic (https://www.kaggle.com/c/titanic))

El dataset forma parte de una competición de Kaggle en el que el objetivo es crear un modelo que sea capaz de predicir las probabilidades de sobrevivir al hundimiento del Titanic en 1912 de los pasajeros, en función de sus características.

En concreto, el dataset de entrenamiento contiene los siguientes campos:

- survival: una variable dicotómica que nos indica si el pasajero sobrevivió o no (0=No, 1=Sí)
- pclass: variable categórica que indica la clase del billete del pasajero (1=primera clase, 2=segunda clase y 3=tercera clase). Se comenta en la descripción que es un proxy (medida indirecta) del status socio-económico del pasajero.
- sex: variable dicotómica con el sexo del pasajero.
- age: variable numérica con la edad en años.
- sibsp: variable numérica con la suma de número de hermanos y esposos a bordo del Titanic.
- parch: variable numérica con la suma del número de padres e hijos a bordo del Titanic.
- ticket: número de ticket.
- fare: variable numérica con el coste del billete.

- cabin: número de cabina.
- embarked: variable categórica que indica en qué ciudad embarcó el pasajero (C = Cherbourg, Q = Queenstown, S = Southampton)

La pregunta que se trata de contestar es: ¿cuáles eran las características de las personas que tenían más probabilidad de sobrevivir? Para ello contamos con los datos de los pasajeros descritos anteriormente. Algunas de las preguntas que se pueden plantear son:

- ¿Tuvieron las mujeres más probabilidades de sobrevivir? Si es así, ¿sucedió en todos los casos o solo cuando viajaban con hijos?
- ¿Tuvieron los pasajeros menor de edad más probabilidades de sobrevivir?
- ¿Qué influencia tuvo la clase del pasaje en la probabilidad de sobrevivir?
- ¿Tiene alguna relación la ciudad de embarque con las probabilidades de sobrevivir? Si es así, ¿se puede correlacionar la ciudad de embarque con otras características: clase del billete, precio, etc.?

El dataset está dividido ya en 2: “train” y “test”. El dataset de “train” es el que se debe usar para entrenar los modelos, mientras que el de “test” se debe usar en la competición para realizar nuestras predicciones y obtener la puntuación del modelo; con lo que no contiene información de si el pasajero sobrevivió o no.

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

Leo los datos, aprovechando para interpretar las cadenas de texto vacías como datos faltantes:

```
titanic_df <- read.csv("data/train.csv", na.strings=c(""))
```

Comprobamos que los datos se han cargado correctamente:

```
head(titanic_df, 10)
```

| | PassengerId | Survived | Pcla… | Name |
|---|---|---|---|---|
| | <int> | <int> | <int> | <fctr> |
| 1 | 1 | 0 | 3 | Braund, Mr. Owen Harris |
| 2 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) |
| 3 | 3 | 1 | 3 | Heikkinen, Miss. Laina |
| 4 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) |
| 5 | 5 | 0 | 3 | Allen, Mr. William Henry |
| 6 | 6 | 0 | 3 | Moran, Mr. James |
| 7 | 7 | 0 | 1 | McCarthy, Mr. Timothy J |
| 8 | 8 | 0 | 3 | Palsson, Master. Gosta Leonard |
| 9 | 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) |
| 10 | 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) |

1-10 of 10 rows | 1-7 of 13 columns

Vemos el tipo de datos de cada columna:

```
sapply(titanic_df, class)
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##   "integer"   "integer"   "integer"    "factor"    "factor"   "numeric"
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##   "integer"   "integer"    "factor"   "numeric"    "factor"    "factor"
```

Vemos que la mayoría de las variables son enteros o númericas. La excepción son “Name”, “Sex”, “Ticket”, “Cabin” y “Embarked”.

La variable “Survived” que fue interpretada como una variable de tipo entero es realmente una variables categórica (dicotómica concretamente), así que procedemos a transformarla a “factor”:

```
titanic_df$Survived_category <- factor(titanic_df$Survived, levels=c(0,1),
                                       labels=c("No", "Yes"))
```

En el caso de la variable “Pclass” fue interpretada como variable numérica, pero puede también considerarse categórica, aunque hay un orden implícito en las categorías, con lo que nos puede interesar mantenerla como variable numérica y además añadirla como variable categórica:

```
titanic_df$Pclass_category <- factor(titanic_df$Pclass, levels=c(1, 2, 3),
                                     labels=c("Primera", "Segunda", "Tercera"))
```

En este punto podemos aprovechar para eliminar algunas variables que no me van a ser útiles en mi análisis (aunque en otros contextos podrían serlo), como “PassengerId” y “Name”:

```
titanic_df <- subset(titanic_df, select = -c(PassengerId, Name) )
```

Usamos la función “summary” para comprobar que las variables son del tipo concreto y hacernos una primera idea de qué valores toman:

```
summary(titanic_df)
```

```
##     Survived          Pclass          Sex           Age            SibSp
##  Min.   :0.0000   Min.   :1.000   female:314   Min.   : 0.42   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:2.000   male  :577   1st Qu.:20.12   1st Qu.:0.000
##  Median :0.0000   Median :3.000                Median :28.00   Median :0.000
##  Mean   :0.3838   Mean   :2.309                Mean   :29.70   Mean   :0.523
##  3rd Qu.:1.0000   3rd Qu.:3.000                3rd Qu.:38.00   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :3.000                Max.   :80.00   Max.   :8.000
##                                                NA's   :177
##      Parch            Ticket         Fare               Cabin     Embarked
##  Min.   :0.0000   1601    :  7   Min.   :  0.00   B96 B98    :  4   C   :168
##  1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91   C23 C25 C27:  4   Q   : 77
##  Median :0.0000   CA. 2343:  7   Median : 14.45   G6         :  4   S   :644
##  Mean   :0.3816   3101295 :  6   Mean   : 32.20   C22 C26    :  3   NA's:  2
##  3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00   D          :  3
##  Max.   :6.0000   CA 2144 :  6   Max.   :512.33   (Other)    :186
##                   (Other) :852                    NA's       :687
##  Survived_category Pclass_category
##  No :549           Primera:216
##  Yes:342           Segunda:184
##                    Tercera:491
##
##
##
##
```

# 3. Limpieza de los datos.

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

En cuanto a elementos vacíos, podemos comprobarlo con la función “is.na” y la función “colSums” para hacer un conteo por columna:

```
colSums(is.na(titanic_df))
```

```
##          Survived            Pclass               Sex               Age
##                 0                 0                 0               177
##             SibSp             Parch            Ticket              Fare
##                 0                 0                 0                 0
##             Cabin          Embarked Survived_category   Pclass_category
##               687                 2                 0                 0
```

Vemos que hay 3 variables que contienen NAs: “Age”, “Cabin” y “Embarked”.

La variable “Age” continee 177 valores vacíos. Como el número de observaciones totales es de 891, la proporción es muy grande; con lo que no podemos prescindir de estas observaciones, es mejor por ejemplo sustitiuir esos Na’s con la media de las edades.

Podríamos usar simplemente la media de los pasajeros, o tener en cuenta que las edades pudo ser bastante diferente en función de otras características. En este caso, para no complicarlo demasiado, vamos a tener en cuenta solo el sexo. Para ello divimos el dataset en mujeres y hombres y calculamos sus edades medias:

```
women_df <-titanic_df[titanic_df$Sex == "female",]
women_mean_age <- mean(women_df$Age, na.rm= TRUE)
print(women_mean_age)
```

```
## [1] 27.91571
```

```
men_df <-titanic_df[titanic_df$Sex == "male",]
men_mean_age <- mean(men_df$Age, na.rm= TRUE)
print(men_mean_age)
```

```
## [1] 30.72664
```

Sólo nos queda aplicarlo a las observaciones que no tienen edad conocida:

```
titanic_df$Age[is.na(titanic_df$Age) & titanic_df$Sex=="female" ] <- women_mean_age
titanic_df$Age[is.na(titanic_df$Age) & titanic_df$Sex=="male" ] <- men_mean_age
```

En cuanto a la variable “Embarked”, solo tenemos 2 observaciones con valores indefinido. Como son tan pocas observaciones, podemos prescindir de ellas, quedándonos con 889 observaciones:

```
titanic_df <- titanic_df[!is.na(titanic_df$Embarked),]
```

Comprobamos que ahora tenemos 2 observaciones menos:

```
nrow(titanic_df)
```

```
## [1] 889
```

En cuanto a la variable “Cabin”, hay una proporción muy grande para la que no tenemos valor (687 observaciones), con lo que no podemos prescindir de estas observaciones. Por otro lado, es una variable que podría darnos algún tipo de información, pero tampoco parece que tenga mucho valor y más cuando en la mayoría de las observaciones este valor no está presente, con lo que podemos simplemente prescindir de esta variable:

```
titanic_df <- subset(titanic_df, select = -c(Cabin))
```

Comprobamos nuevamente el número de NA’s por columna:

```
colSums(is.na(titanic_df))
```

```
##          Survived            Pclass               Sex               Age 
##                 0                 0                 0                 0 
##             SibSp             Parch            Ticket              Fare 
##                 0                 0                 0                 0 
##          Embarked Survived_category   Pclass_category 
##                 0                 0                 0
```

Confirmamos que ya no quedan columnas con NA’s.

Tras comprobar y tratar los NA’s, tenemos que mirar si hay ceros que en realidad identifican falta de datos. Podemos usar la función “min” para ver los valores mínimos de las variables numéricas:

```
sapply(titanic_df[c("Age", "SibSp", "Parch", "Fare")], min)
```

```
##   Age SibSp Parch  Fare
##  0.42  0.00  0.00  0.00
```

Vemos que en el caso de “Age” el valor mínimo es 0.42; es un valor correcto porque para niños de menos de un año, la edad está definida como una fracción. En el caso de “SibSp” y “Parch”, que son el número de hermanos/cónyuges y el número de padres y hijos, el valor 0 es totalmente normal.

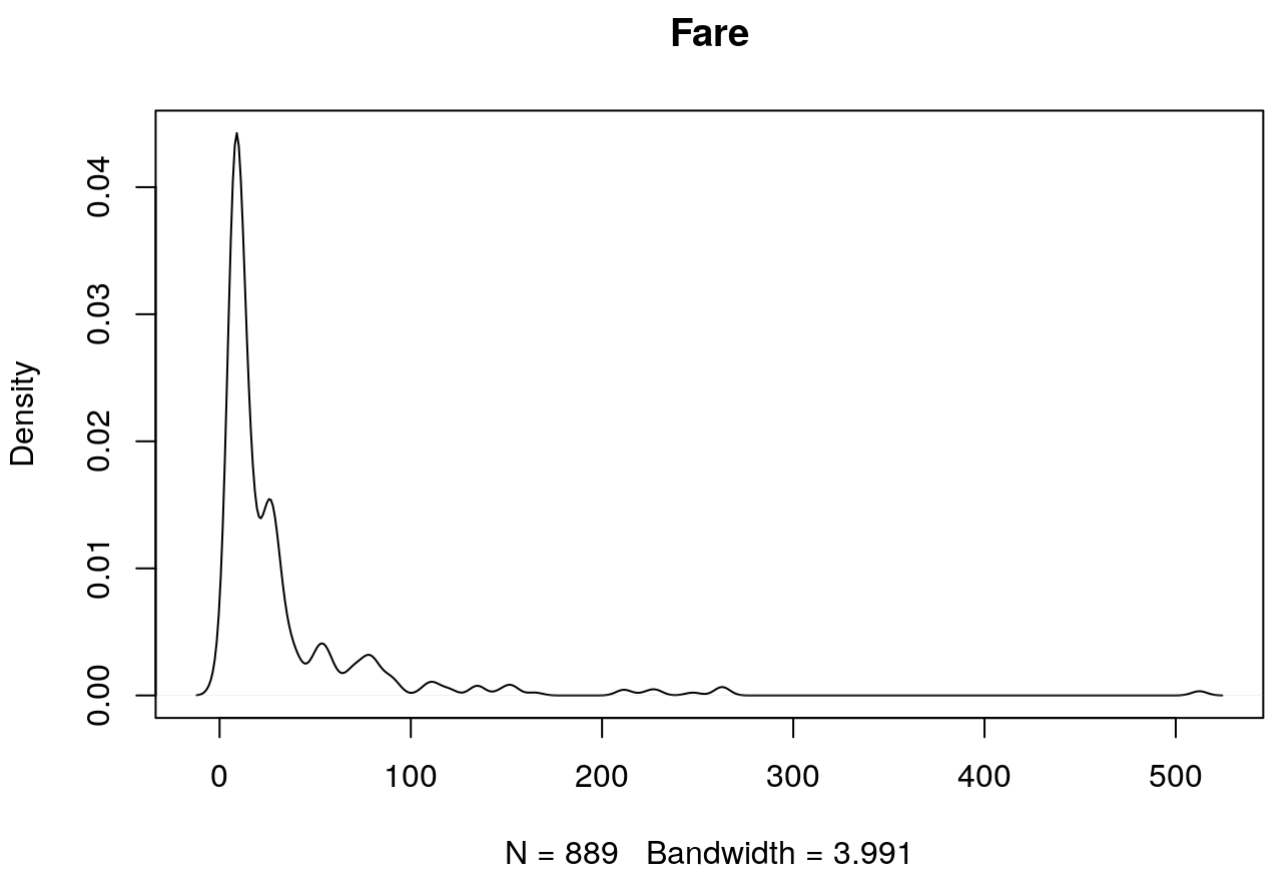
En el caso de “Fare” (coste del billete) no tiene mucho sentido tener ceros (aunque podría darse si hubiese invitaciones, pero vamos a asumir que esto no es posible). Vamos a comprobar cuantos ceros hay:

```
nrow(titanic_df[titanic_df$Fare == 0,])
```

```
## [1] 15
```

Tenemos 15 observaciones con un valor para “Fare” de 0. Podemos hacer una gráfica de densidades para ver como se distribuyen los valores:

```
plot(density(titanic_df$Fare), main="Fare")
```




Vemos visualmente que la mayor parte de los valores se encuentran entre 0 y 50. Más específicamente podemos ver usando “summary” que la media es de 32.097 y la mediana es de 14.454:

```
summary(titanic_df$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.896  14.454  32.097  31.000 512.329
```

Los valores con 0 están relativamente lejos de la media y la mediana, así que es mejor que tratemos esos valores. Vemos también que la media está desplazada debido a un outlier (como veremos en el siguiente apartado), podríamos ignorarlo o podemos simplemente usar la mediana.

Además podemos ver que la mediana del precio (“Fare”) es muy diferente en función de la clase del pasaje:

```
median_primera = median(titanic_df$Fare[titanic_df$Pclass_category == "Primera"])
print(median_primera)
```

```
## [1] 58.6896
```

```
median_segunda = median(titanic_df$Fare[titanic_df$Pclass_category == "Segunda"])
print(median_segunda)
```

```
## [1] 14.25
```

```
median_tercera = median(titanic_df$Fare[titanic_df$Pclass_category == "Tercera"])
print(median_tercera)
```

```
## [1] 8.05
```

Podríamos tener en cuenta más factores, o podríamos crear un modelo de regresión lineal para tratar esos valores, pero como son sólo 15, considero que es suficiente con usar la mediana en función de la clase del pasaje:

```
titanic_df$Fare[titanic_df$Fare == 0 & titanic_df$Pclass_category=="Primera" ] <- med
ian_primera
titanic_df$Fare[titanic_df$Fare == 0 & titanic_df$Pclass_category=="Segunda" ] <- med
ian_segunda
titanic_df$Fare[titanic_df$Fare == 0 & titanic_df$Pclass_category=="Tercera" ] <- med
ian_tercera
```

Comprobamos que la variable “Fare” ya no tiene ceros:

```
summary(titanic_df$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.013   7.925  14.458  32.559  31.275 512.329
```

## 3.2. Identificación y tratamiento de valores extremos.

Las variabes numéricas que tenemos son “Age”, “SibSp”, “Parch” y “Fare”. Podemos ver algunos de sus características usando “summary”:

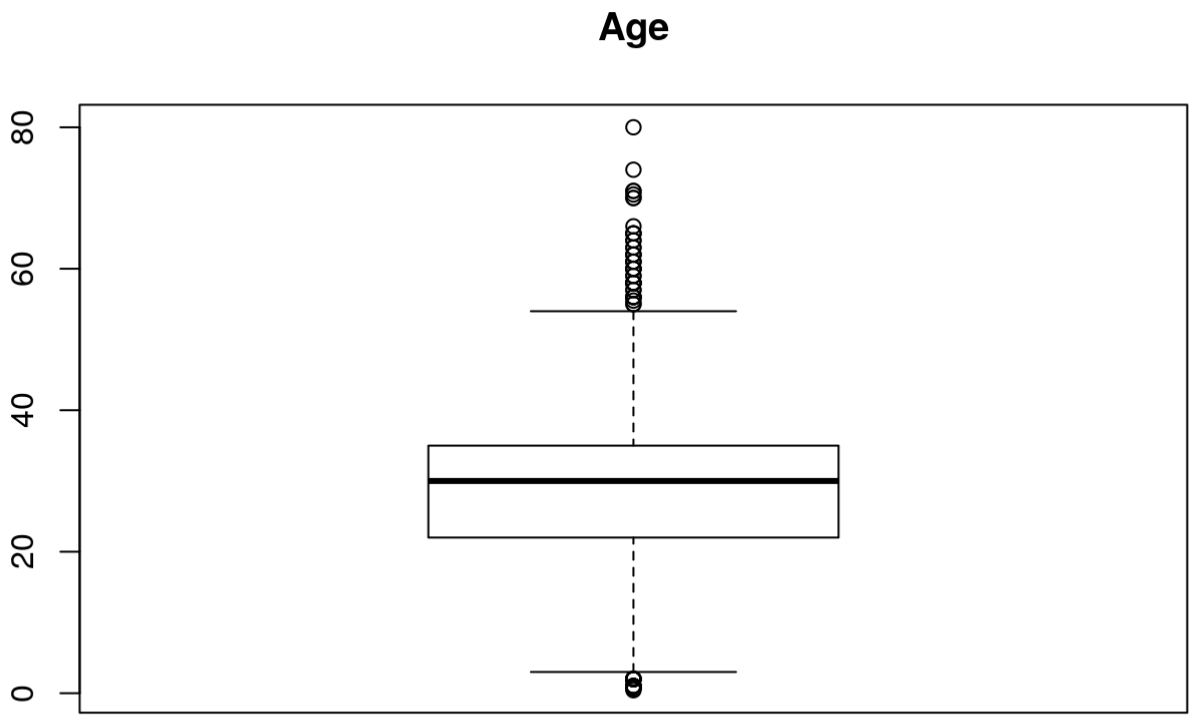
```
summary(titanic_df[c("Age", "SibSp", "Parch", "Fare")])
```

```
##       Age            SibSp            Parch            Fare
##  Min.   : 0.42   Min.   :0.0000   Min.   :0.0000   Min.   :  4.013
##  1st Qu.:22.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  7.925
##  Median :30.00   Median :0.0000   Median :0.0000   Median : 14.458
##  Mean   :29.69   Mean   :0.5242   Mean   :0.3825   Mean   : 32.559
##  3rd Qu.:35.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.: 31.275
##  Max.   :80.00   Max.   :8.0000   Max.   :6.0000   Max.   :512.329
```

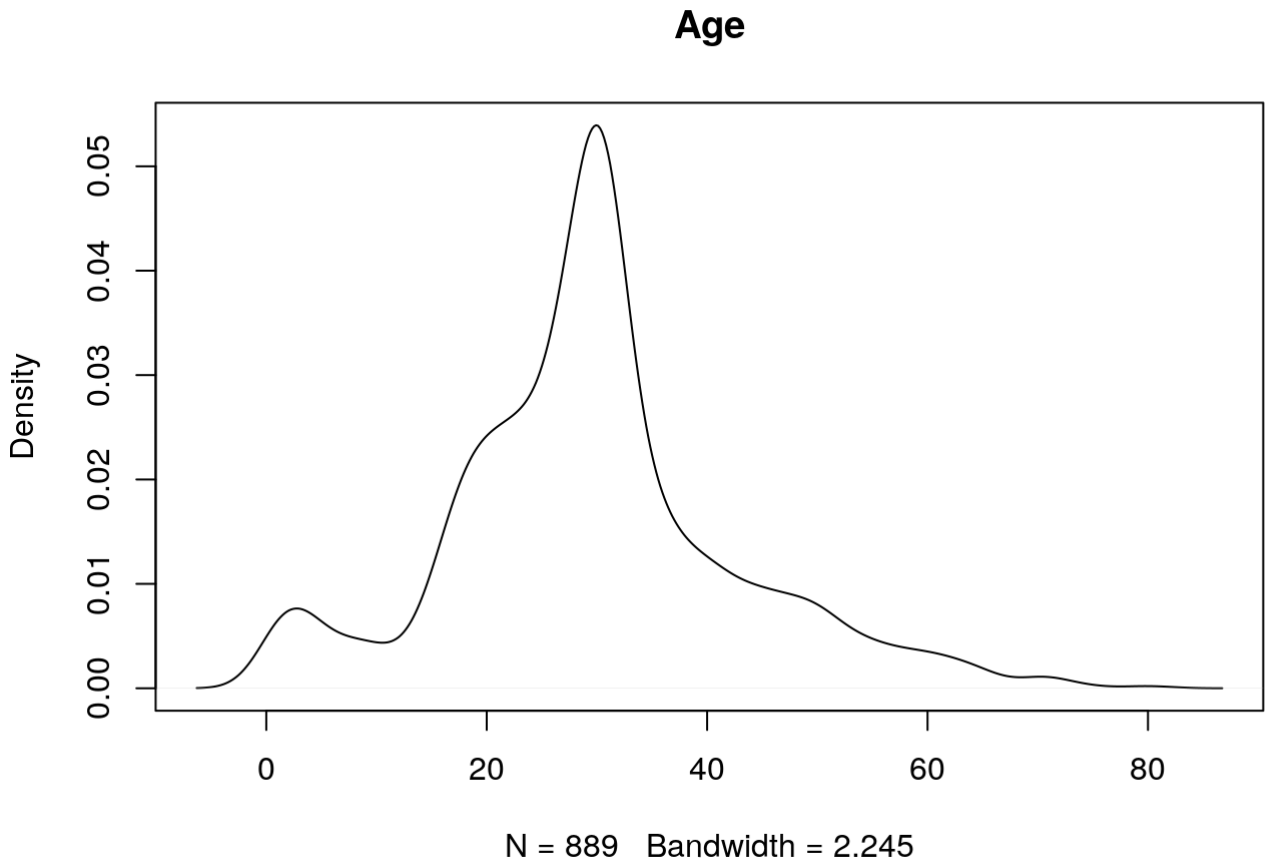
Vemos que los valores de “Age” se encuentran entre 0.42 y 80, con una mediana de 30 y una media de 29.69, parece totalmente correcto. En el caso de “SibSp” y “Parch”, que es el número de hermanos/conyúges y padres/hijos, los valores se encuentran entre 0-8 y 0-6 respectivamente, que también parece correcto. En el caso de “Fare” el rango es mayor, entre 4.013 y 512.329, con una mediana de 14.458 y una media de 32.559; lo que nos hace pensar que hay outliers.

Para ver si hay outliers, usamos “boxplot” y gráficas de densidad (para ver como se distribuyen los valores), con cada una de estas variables:

```
boxplot(titanic_df$Age, main="Age")
```

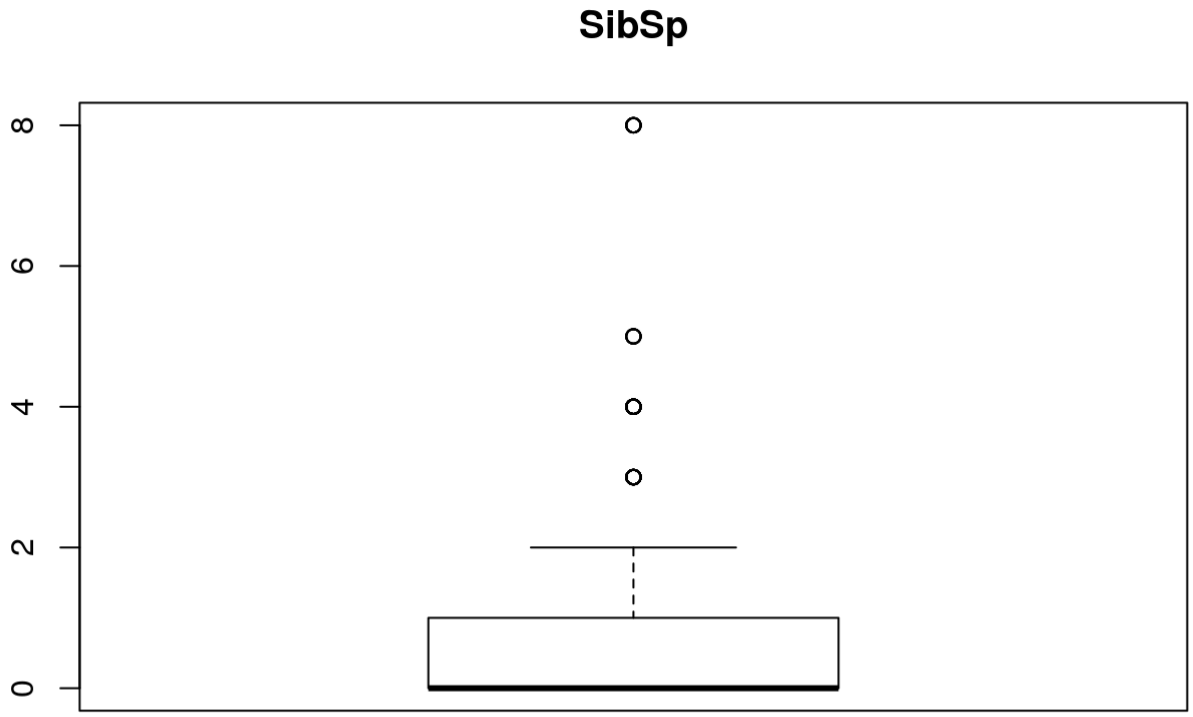



```
plot(density(titanic_df$Age), main="Age")
```

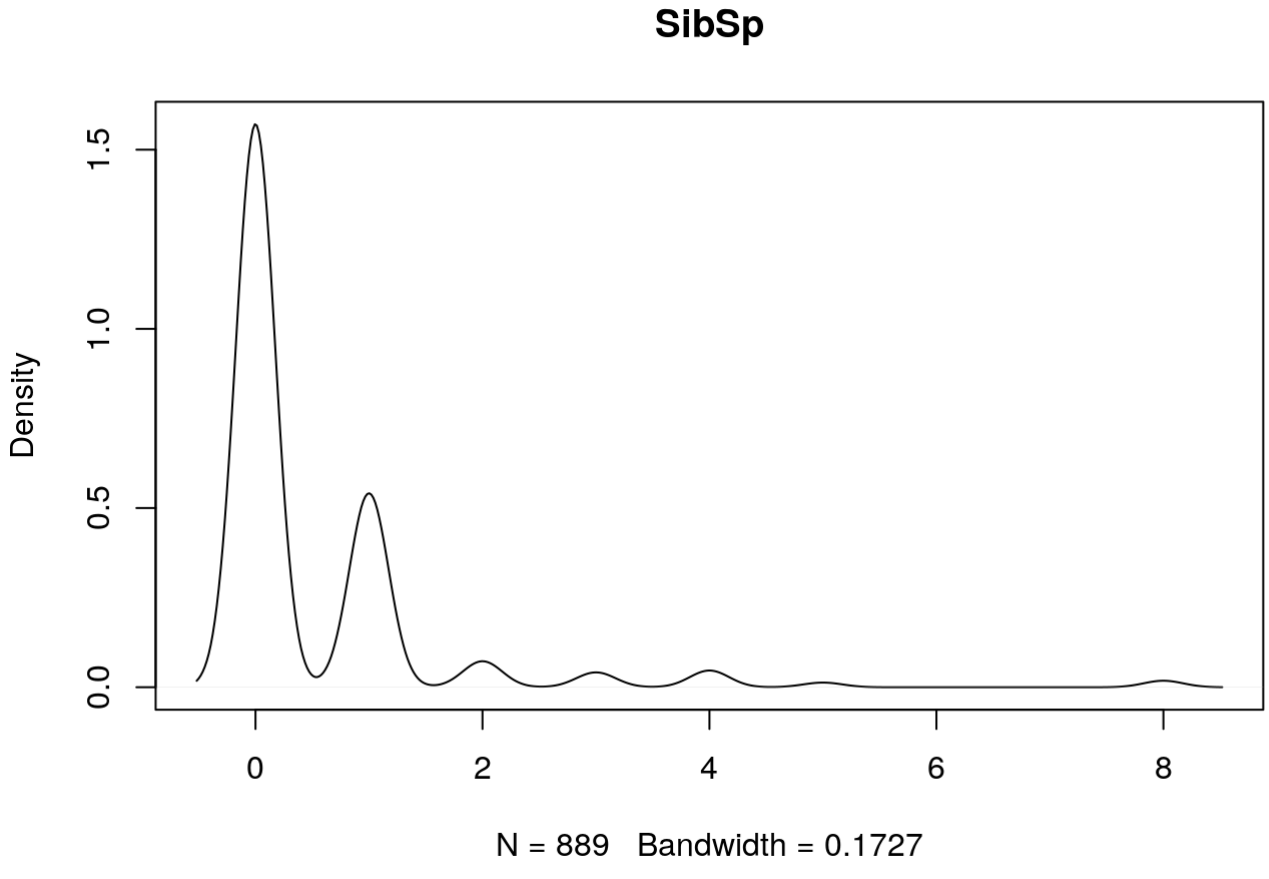



Vemos que la variable “Age” contiene “outliers”, pero son valores correctos, que nos dan información útil, con lo que no necesitamos tratarlos (eliminarlos sería contraproducente).

```
boxplot(titanic_df$SibSp, main="SibSp")
```

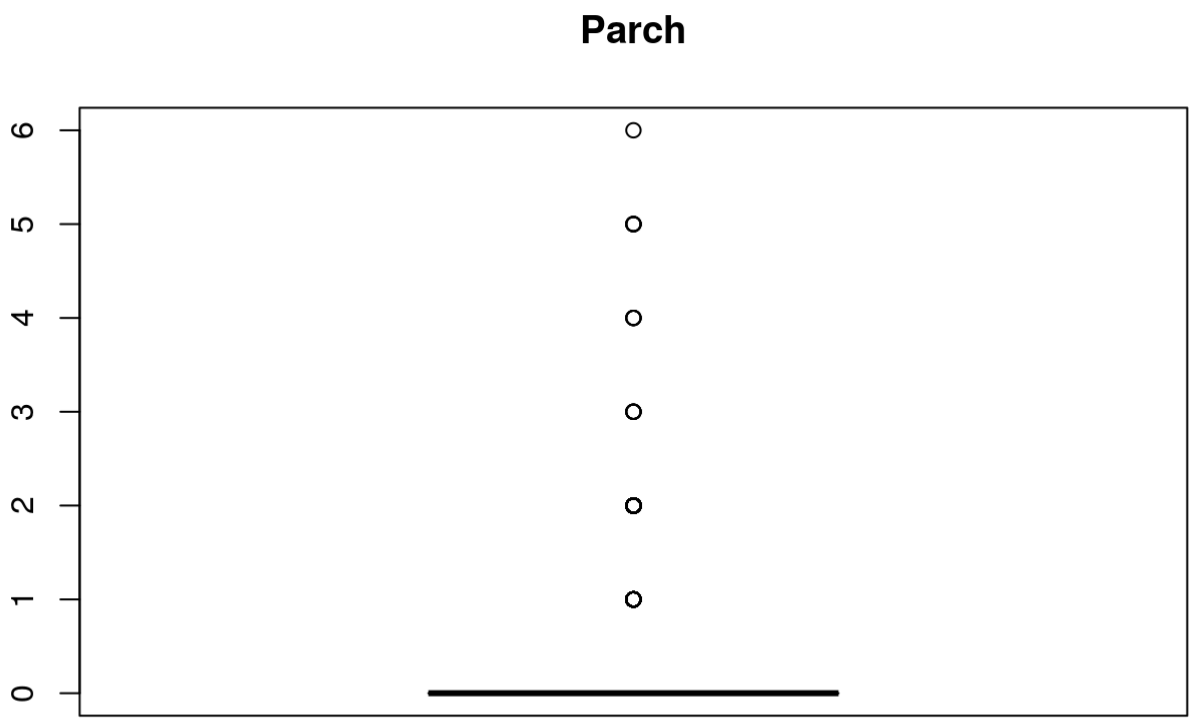



```
plot(density(titanic_df$SibSp), main="SibSp")
```

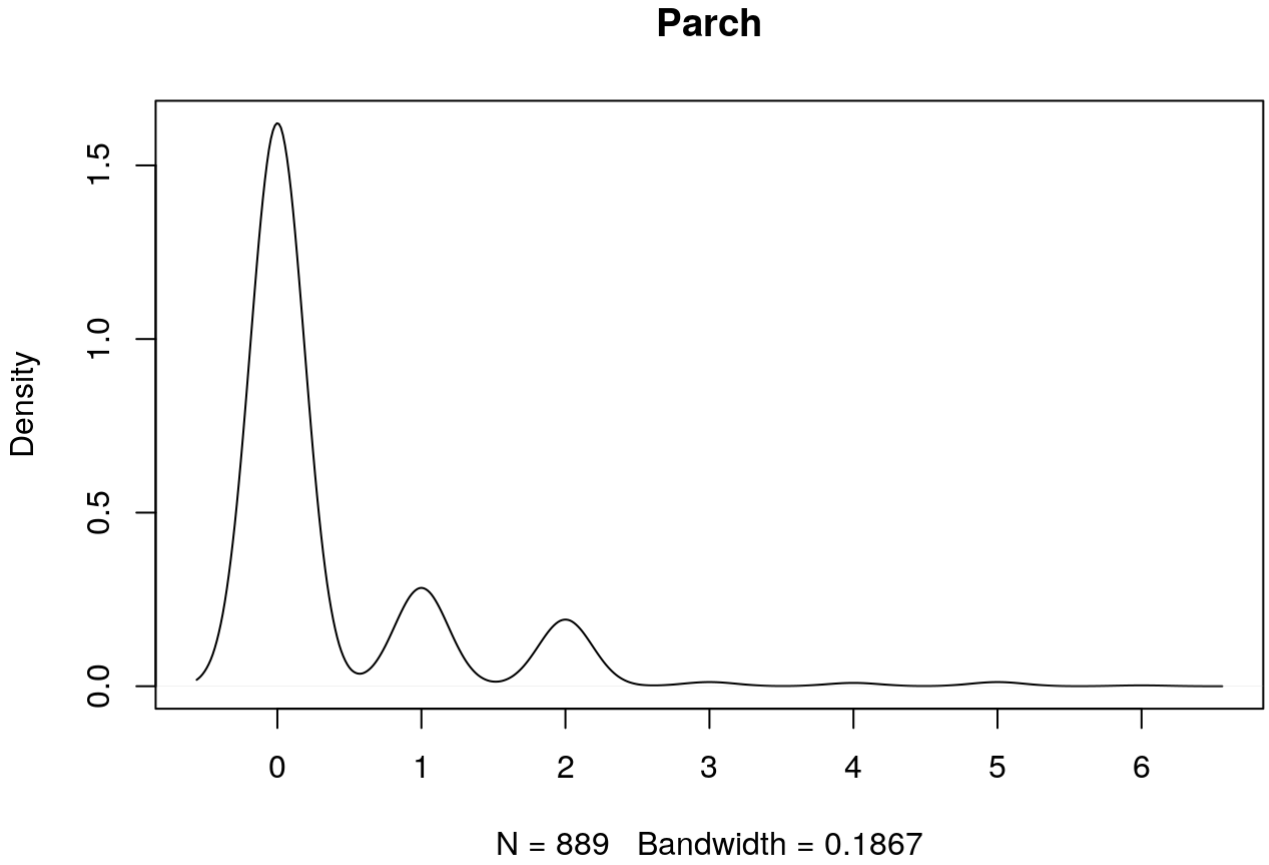


Vemos que la variable “SibSp” también contiene “outliers” (son los valores 3, 4, 5 y 8), pero son también valores correctos y son pocos, así que considero que es mejor no tratarlos.

```
boxplot(titanic_df$Parch, main="Parch")
```

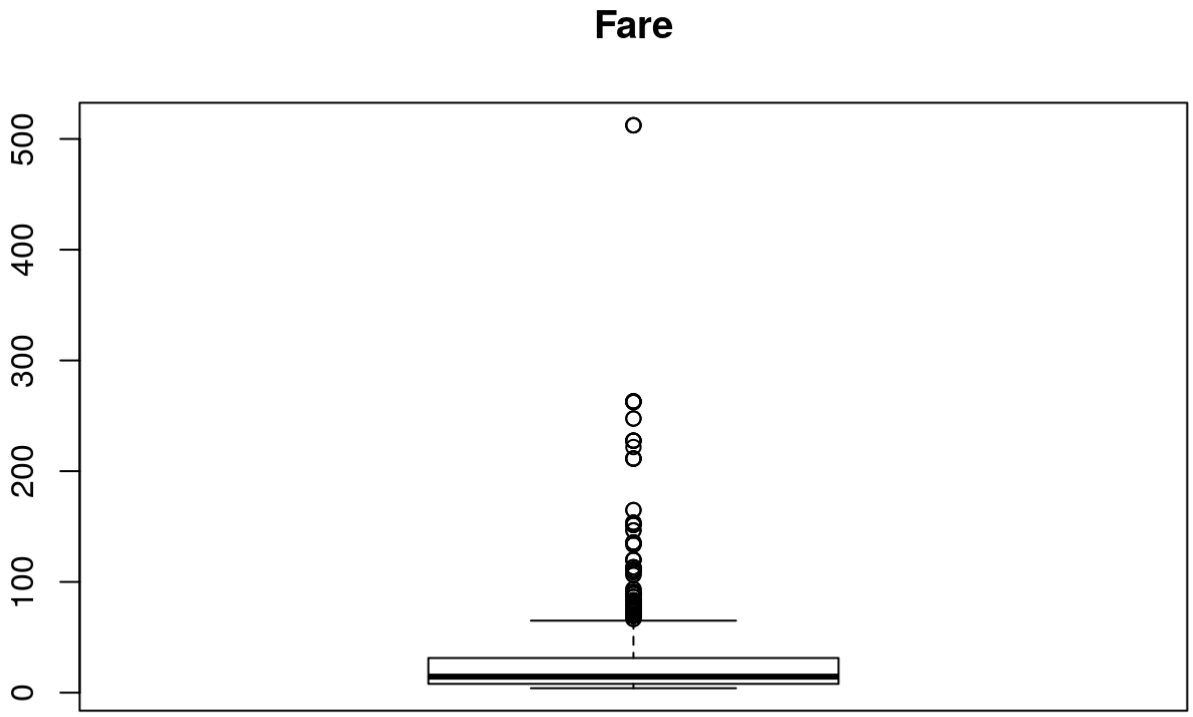



```
plot(density(titanic_df$Parch), main="Parch")
```

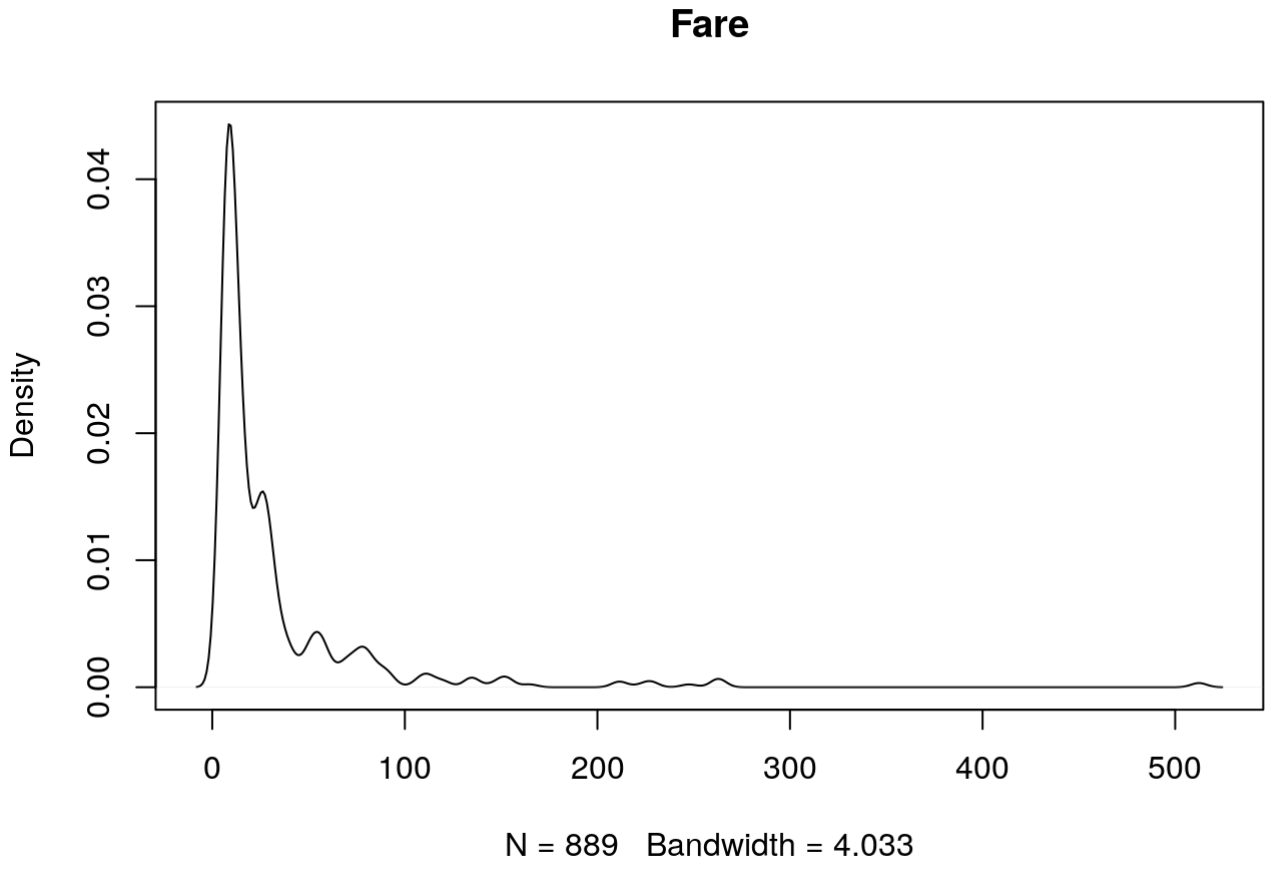



Es un caso similar al de “Parch”. La mayor parte de los valores son 0, de hecho la mediana es 0. Algunos valores serían por definición “outliers”, pero son valores correctos y útiles para nuestro análisis.

```
boxplot(titanic_df$Fare, main="Fare")
```




```
plot(density(titanic_df$Fare), main="Fare")
```



En el caso de la variable “Fare”, tenemos bastantes “outliers”, pero la mayoría están cercanos entre ellos. Sin embargo, hay unos valores muy alejados, mayores a 500; cuando la mediana es 14.458 y la media es 32.559. Vamos a analizar en detalle estas observaciones:

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

| | Survived <int> | Pclass <int> | Sex <fctr> | … <dbl> | SibSp <int> | Parch <int> | Ticket <fctr> | Fare <dbl> | Embarked <fctr> |
|---|---|---|---|---|---|---|---|---|---|
| 259 | 1 | 1 | female | 35 | 0 | 0 | PC 17755 | 512.3292 | C |
| 680 | 1 | 1 | male | 36 | 0 | 1 | PC 17755 | 512.3292 | C |
| 738 | 1 | 1 | male | 35 | 0 | 0 | PC 17755 | 512.3292 | C |

3 rows | 1-10 of 12 columns

Vemos que son 3 observaciones, dónde coincide exactamente el coste (512.33), la clase (“Primera”), el ticket (“PC 17755”) y el lugar dónde embarcaron. Parece por tanto que son valores correctos. Sin embargo, aun teniendo en cuenta que son pasajes de “Primera” están muy alejados de la media para esta clase, estos valores desplazan le media y usarlos prodría ser contraproducente para nuestros modelos. Aunque hay que decir que en este caso, las 3 observaciones corresponden a pasajeros que sobrevivieron, y forman parte de la “Primera” clase, siendo los billetes más caros, que “a priori” podemos pensar que son los pasajeros con más probabilidad de sobrevivir.

Vamos a mantenerlos de momento, pero lo tenemos presente a la hora de crear nuestros modelos por si fuese mejor no tenerlos en cuenta.

# 4. Análisis de los datos.

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

Los grupos que vamos a analizar/comparar van a ser en función de las variables Sex, Pclass_category y Embarked:

```
women_titanic <- titanic_df[titanic_df$Sex == "female",]
men_titanic <- titanic_df[titanic_df$Sex == "male",]

primera_titanic <- titanic_df[titanic_df$Pclass_category == "Primera",]
segunda_titanic <- titanic_df[titanic_df$Pclass_category == "Segunda",]
tercera_titanic <- titanic_df[titanic_df$Pclass_category == "Tercera",]

cherbourg_titanic <- titanic_df[titanic_df$Embarked == "C",]
queenstown_titanic <- titanic_df[titanic_df$Embarked == "Q",]
southampton_titanic <- titanic_df[titanic_df$Embarked == "S",]
```

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

Para comprobar la normalidad de nuestras variables numéricas podemos usar el test de Anderson-Darling:

```
library(nortest)
age_ad_pvalue <- ad.test(titanic_df$Age)$p.value
print(c("Age Anderson-Darling p_value:", age_ad_pvalue))
```

```
## [1] "Age Anderson-Darling p_value:" "3.7e-24"
```

```
sibsp_ad_pvalue <- ad.test(titanic_df$SibSp)$p.value
print(c("Sibsp Anderson-Darling p_value:", sibsp_ad_pvalue))
```

```
## [1] "Sibsp Anderson-Darling p_value:" "3.7e-24"
```

```
parch_ad_pvalue <- ad.test(titanic_df$Parch)$p.value
print(c("Parch Anderson-Darling p_value:", parch_ad_pvalue))
```

```
## [1] "Parch Anderson-Darling p_value:" "3.7e-24"
```

```
fare_ad_pvalue <- ad.test(titanic_df$Fare)$p.value
print(c("Fare Anderson-Darling p_value:", fare_ad_pvalue))
```

```
## [1] "Fare Anderson-Darling p_value:" "3.7e-24"
```

En todos los casos el p_value es muy pequeño Para un nivel de significancia de 0.05, que se suele usar comúnmente, tenemos que rechazar la hipótesis nula; por tanto, las variables no siguen una distribución normal.

Para analizar la homogeneidad de las varianzas usaremos el test de Fligner-Killeen. Para ello tenemos que indicar la variable cuya varianza se va a testear y la variable que se va a usar para formar los grupos. Como comentamos vamos a usar la clase del pasaje, el sexo y dónde embarcaron:

```
fligner.test(Survived ~ Pclass_category, data = titanic_df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Pclass_category
## Fligner-Killeen:med chi-squared = 36.046, df = 2, p-value = 1.488e-08
```

```
fligner.test(Survived ~ Sex, data = titanic_df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Sex
## Fligner-Killeen:med chi-squared = 6.0178, df = 1, p-value = 0.01416
```

```
fligner.test(Survived ~ Embarked, data = titanic_df)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Survived by Embarked
## Fligner-Killeen:med chi-squared = 7.1781, df = 2, p-value = 0.02762
```

En los 3 casos el valor de p_value es inferior al nivel de significancia 0.05, por tanto, tenemos que rechazar la hipótesis nula y, por tanto, para los grupos que hemos analizado la varianza de “Survived” no es homogénea. Esto nos da una indicación de que las variables que analizamos (clase, sexo y lugar de embarque) van a ser importantes a la hora de predecir si un pasajero sobrevivió o no.

## 4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos.

### 4.3.1 Correlaciones

Empezamos estudiando las correlaciones de las variables numéricas con respecto a la variable “Survived” usando el coeficiente de correlación de Spearman:

```
cor_pclass <- cor.test(titanic_df$Survived, titanic_df$Pclass, method = "spearman")
```

```
## Warning in cor.test.default(titanic_df$Survived, titanic_df$Pclass, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
cor_age <- cor.test(titanic_df$Survived, titanic_df$Age, method = "spearman")
```

```
## Warning in cor.test.default(titanic_df$Survived, titanic_df$Age, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
cor_sibsp <- cor.test(titanic_df$Survived, titanic_df$SibSp, method = "spearman")
```

```
## Warning in cor.test.default(titanic_df$Survived, titanic_df$SibSp, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
cor_parch <- cor.test(titanic_df$Survived, titanic_df$Parch, method = "spearman")
```

```
## Warning in cor.test.default(titanic_df$Survived, titanic_df$Parch, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
cor_fare <- cor.test(titanic_df$Survived, titanic_df$Fare, method = "spearman")
```

```
## Warning in cor.test.default(titanic_df$Survived, titanic_df$Fare, method =
## "spearman"): Cannot compute exact p-value with ties
```

Presentamos los resultados en una tabla:

```
library(kableExtra)
out <- data.frame(variable=c("Pclass", "Age", "SibSp", "Parch", "Fare"),
                  estimate=c(cor_pclass$estimate, cor_age$estimate,
                             cor_sibsp$estimate, cor_parch$estimate,
                             cor_fare$estimate),
                  p_value=c(cor_pclass$p.value, cor_age$p.value,
                            cor_sibsp$p.value, cor_parch$p.value,
                            cor_fare$p.value)
                  )
out %>% kable() %>% kable_styling()
```

| variable | estimate | p_value |
|---|---|---|
| Pclass | -0.3369167 | 0.0000000 |
| Age | -0.0747634 | 0.0258047 |
| SibSp | 0.0909440 | 0.0066596 |
| Parch | 0.1401263 | 0.0000275 |
| Fare | 0.3034107 | 0.0000000 |

Vemos que la correlación más fuerte se produce con “Pclass” y es de signo negativo. Es decir, los pasajeros con Pclass=1, que son los de “Primera” clase, serían los que tienen más probabilidades de sobrevivir. A medida que el valor numérico de Pclass sube, las probabilidades de sobrevivir bajan.

La siguiente con más fuerza sería “Fare”, de signo positivo. Es decir, contra más caro es el billete, más probabilidades de sobrevivir se tienen. Esta variable se correlaciona negativamente con “Pclass”, así que este resultado tiene sentido.

En el caso de “Parch” y “SibSp” vemos correlaciones con poco fuerza, pero que podrían indicar que los nucleos familiares tuvieron más posibilidades de sobrevivir que los que viajaban solos.

Por último, la variable “Age” correlaciona negativamente con “Survived”, pero sin mucha fuerza. Esto indicaría que los niños y jóvenes tuvieron más posibilidad de sobrevivir, aunque el coeficiente de correlación y el p_value nos indica que no tiene mucha fuerza.

### 4.3.2 Contrastes de hipótesis

A continuación, realizaremos unos contrastes de hipótesis entre los diferentes grupos que definimos anteriormente; en concreto, usaremos un “contraste de hipótesis de dos muestras independientes sobre la proporción”. Es decir, vamos a comprobar si las diferencias en la proporción de supervivientes entre diferentes grupos es significativa o no.

Al ser la muestra mayor de 30 observaciones, en aplicación del Teorema del Límite Central, podemos asumir normalidad.

Definimos la hipótesis nula y alternativa para el caso en el que comparamos hombres y mujeres:

$$H_0 : p(women) = p(men)$$

$$H_1 : p(women) > p(men)$$

Defino la hipótesis alternativa como que la proporción de supervivientes en el grupo de mujeres es mayor que en el grupo de hombres, así que es un test unilateral.

```
n_women <- nrow(women_titanic)
n_men <- nrow(men_titanic)

survivors_women <- sum(women_titanic$Survived == 1)
survivors_men <- sum(men_titanic$Survived == 1)

survivors <-c (survivors_women, survivors_men)
nn <- c(n_women, n_men)
prop.test(survivors, nn, correct=FALSE, alternative="greater", conf.level=0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  survivors out of nn
## X-squared = 260.76, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.5026372 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.7403846 0.1889081
```

Vemos que para un nivel de confianza del 95%, obtenemos un p_value muy por debajo del valor de significancia (0.05). Podemos decir que la diferencia en la proporción de supervivientes entre las mujeres es significativo y, por tanto, las mujeres tenían más probabilidades de sobrevivir.

A continuación realizaremos el contraste comparando los pasajeros por la clase en la que viajaban. Compararemos primero los pasajeros de “Primera” con respecto a los demás y, a continuación, haremos lo mismo pero solo teniendo en cuenta a hombres (más adelante veremos por qué).

Definimos la hipótesis nula y alternativa para el primer caso. La hipótesis alternativa que voy a usar es que la proporción de supervivientes en “Primera” es mayor, por tanto, es un test unilateral:

$$H_0 : p(primera) = p(resto)$$

$$H_1 : p(primera) > p(resto)$$

```
primera_titanic <- titanic_df[titanic_df$Pclass_category == "Primera",]
resto_titanic <- titanic_df[titanic_df$Pclass_category != "Primera",]

n_primera <- nrow(primera_titanic)
n_resto <- nrow(resto_titanic)

survivors_primera <- sum(primera_titanic$Survived == 1)
survivors_resto <- sum(resto_titanic$Survived == 1)

survivors <-c (survivors_primera, survivors_resto)
nn <- c(n_primera, n_resto)
prop.test(survivors, nn, correct=FALSE, alternative="greater", conf.level=0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  survivors out of nn
## X-squared = 70.881, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.259263 1.000000
## sample estimates:
##    prop 1    prop 2
## 0.6261682 0.3051852
```

Vemos en el resultado que efectivamente la proporción de supervivientes en Primera es de más de un 62% mientras que en el resto es de un 30%; y que además esta diferencia es significativa ya que el “p_value” es menor a 0.05. Confirmamos, por tanto, que los pasajeros de “Primera” tuvieron más probabilidades de sobrevivir.

Podemos crear una tabla de frecuencias relativas para comprobar que la proporción de mujeres en “Primera” y “Segunda” era mayor que en “Tercera” (en la que más de un 70% eran hombres):

```
freq_class <- table(titanic_df$Pclass, titanic_df$Sex)
prop.table(freq_class, 1)
```

```
##
##        female      male
##   1 0.4299065 0.5700935
##   2 0.4130435 0.5869565
##   3 0.2932790 0.7067210
```

Como en el contraste anterior obtuvimos que las mujeres tenían más probabilidades de sobrevivir, esto ha podido influir en el resultado. Vamos a realizar el mismo contraste pero solo para hombres, para ver si un hombre que viajase en “Primera” tuvo realmente más probabilidades de sobrevivir que un hombre viajando en otra clase. Las hipótesis nula y alternativa son:

$$H_0 : p(men.primera) = p(men.resto)$$

$$H_1 : p(men.primera) > p(men.resto)$$

```
men_primera_titanic <- men_titanic[men_titanic$Pclass_category == "Primera",]
men_resto_titanic <- men_titanic[men_titanic$Pclass_category != "Primera",]

n_men_primera <- nrow(men_primera_titanic)
n_men_resto <- nrow(men_resto_titanic)

survivors_men_primera <- sum(men_primera_titanic$Survived == 1)
survivors_men_resto <- sum(men_resto_titanic$Survived == 1)

survivors <-c (survivors_men_primera, survivors_men_resto)
nn <- c(n_men_primera, n_men_resto)
prop.test(survivors, nn, correct=FALSE, alternative="greater", conf.level=0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  survivors out of nn
## X-squared = 32.695, df = 1, p-value = 5.391e-09
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.1515024 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.3688525 0.1406593
```

Vemos que aunque las proporciones son en este caso más bajas (por el efecto de no tener en cuenta las mujeres), la diferencia en las proporciones es todavía mayor. El p_value por debajo de 0.05 nos confirma que los hombres que viajaban en “Primera” tenían muchas más probabilidades de sobrevivir que el resto de

hombres.

Por último, realizaremos un contraste teniendo en cuenta la ciudad de embarque. En concreto, queremos saber si los pasajeros que embarcaron en “Cherbourg” tienen una proporción de supervivencia diferente al resto de pasajeros. Es un test bilateral con las siguientes hipótesis:

$$H_0 : p(cherbourg) = p(others)$$

$$H_1 : p(cherbourg)! = p(others)$$

```
cherbourg_titanic <- titanic_df[titanic_df$Embarked == "C",]
others_titanic <- titanic_df[titanic_df$Embarked != "C",]

n_cherbourg <- nrow(cherbourg_titanic)
n_others <- nrow(others_titanic)

survivors_cherbourg <- sum(cherbourg_titanic$Survived == 1)
survivors_others <- sum(others_titanic$Survived == 1)

survivors <-c (survivors_cherbourg, survivors_others)
nn <- c(n_cherbourg, n_others)
prop.test(survivors, nn, correct=FALSE, alternative="two.sided", conf.level=0.95)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  survivors out of nn
## X-squared = 25.682, df = 1, p-value = 4.026e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1282222 0.2937612
## sample estimates:
##    prop 1    prop 2
## 0.5535714 0.3425798
```

El “p_value” es menor a 0.05, con lo que el resultado nos muestra que efectivamente hay una diferencia en la proporción de supervivientes que embarcaron en “Cherbourg” con respecto a los demás. Vemos además que la proporción de supervivientes entre los pasajeros que embarcaron en esta ciudad es de más de un 55%, mientras que para el resto es de un 34%.

Sin embargo, es bastante obvio que la causa de que sobrevivirian más no tiene que ver con la ciudad de embarque en sí, si no que ha de haber otras causas subyacentes. Con la información de los contrastes que hemos hecho anteriormente, sabemos que las mujeres y los pasajeros de “Primera” tenían más probabilidades de sobrevivir, vamos a ver la proporción de estos entre los pasajeros embarcados en las diferentes ciudades:

```
freq_sex <- table(titanic_df$Sex, titanic_df$Embarked)
prop.table(freq_sex, 2)
```

```
##
##                  C         Q         S
##   female 0.4345238 0.4675325 0.3152174
##   male   0.5654762 0.5324675 0.6847826
```

```
freq_class <- table(titanic_df$Pclass, titanic_df$Embarked)
prop.table(freq_class, 2)
```

```
##
##              C          Q          S
##   1 0.50595238 0.02597403 0.19720497
##   2 0.10119048 0.03896104 0.25465839
##   3 0.39285714 0.93506494 0.54813665
```

Vemos que la proporción de mujeres es parecida entre Cherbourg y Queenstown; sin embargo, la proporción de viajeros que embarcaron en Cherbourg en “Primera” clase es de más de un 50%, mucho mayor que en las otras ciudades (de hecho en Queenstown un 93.5% embarcaron en “Tercera” clase, y en Southampton menos de un 20% embarcaron en “Primera”). Por supuesto, puede haber más causas que no estamos analizando aquí, pero estas prodrían explicar en gran medida el resultado.

### 4.3.3 Regresión logística

Por último vamos a entrenar un modelo de tipo regresión logística, ya que la variable que queremos predecir es de tipo dicotómica. Empezamos usando todas las variables que hemos analizado:

```
logistic_model_all = glm(Survived~Sex + Age + SibSp + Parch + Fare + Embarked + Pclas
s_category, family=binomial(), data=titanic_df)
summary(logistic_model_all)
```

```
##
## Call:
## glm(formula = Survived ~ Sex + Age + SibSp + Parch + Fare + Embarked +
##     Pclass_category, family = binomial(), data = titanic_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6108  -0.6117  -0.4178   0.6230   2.4557
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              4.180681   0.477743   8.751  < 2e-16 ***
## Sexmale                 -2.693287   0.201000 -13.399  < 2e-16 ***
## Age                     -0.040051   0.007912  -5.062 4.14e-07 ***
## SibSp                   -0.321151   0.109277  -2.939  0.00329 **
## Parch                   -0.085379   0.118905  -0.718  0.47273
## Fare                     0.001465   0.002395   0.612  0.54063
## EmbarkedQ               -0.053486   0.383205  -0.140  0.88900
## EmbarkedS               -0.456249   0.239348  -1.906  0.05662 .
## Pclass_categorySegunda -0.967672   0.299372  -3.232  0.00123 **
## Pclass_categoryTercera -2.205663   0.300312  -7.345 2.06e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  783.58  on 879  degrees of freedom
## AIC: 803.58
##
## Number of Fisher Scoring iterations: 5
```

El valor AIC del modelo es de 803.58, que es una medida que nos permitirá comparar este modelo con los siguientes.

Vemos que hay algunas variables que no son significativas, ya que tienen un valor Pr(>|z|) mayor a 0,05. Es el caso de “Parch” y “Fare” (en este último caso seguramente porque “Pclass_category” correlaciones fuertemente con ella)· En el caso de “Embarked”, vemos que ninguna de las variables “dummy” son realmente significativas (seguramente por el mismo motivo; por sí sola esta variable da información, pero es un “proxy” de las variable “Pclass_category” y “Sex” como ya vimos), con lo que también la podemos eliminar.

```
logistic_model_optimized = glm(Survived~Sex + Age + SibSp + Pclass_category, family=b
inomial(), data=titanic_df)
summary(logistic_model_optimized)
```

```
##
## Call:
## glm(formula = Survived ~ Sex + Age + SibSp + Pclass_category,
##     family = binomial(), data = titanic_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6936  -0.6036  -0.4218   0.6181   2.4705
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              4.043164   0.399749  10.114  < 2e-16 ***
## Sexmale                 -2.713681   0.194116 -13.980  < 2e-16 ***
## Age                     -0.040851   0.007844  -5.208 1.91e-07 ***
## SibSp                   -0.360642   0.104048  -3.466 0.000528 ***
## Pclass_categorySegunda -1.190375   0.261943  -4.544 5.51e-06 ***
## Pclass_categoryTercera -2.356145   0.243950  -9.658  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  789.41  on 883  degrees of freedom
## AIC: 801.41
##
## Number of Fisher Scoring iterations: 5
```

Vemos que el AIC de este modelo es 801.41, es por tanto mejor que el anterior. Vemos que captura un poco menos de la varianza residual, pero la diferencia es muy poca, pero a cambio tenemos un modelo menos complejo y más eficiente.

Calculamos los OR:

```
exp(coefficients(logistic_model_optimized))
```

```
##           (Intercept)               Sexmale                   Age
##           57.00644256            0.06629233            0.95997218
##                 SibSp Pclass_categorySegunda Pclass_categoryTercera
##            0.69722881            0.30410706            0.09478492
```

Vemos que siendo hombre la probabilidad de sobrevivir bajan un 93.4% respecto a ser mujer. Si viajas en “Segunda” tienes un 70% menos probabilidad de sobrevivir respecto a ir en “Primera”; pero si viajas en “Tercera”, las probabilidades bajan un 90%. A medida que la edad sube, tus posibilidades de sobrevivir bajan.

Podemos comparar varios casos para ver cómo se comporta el modelo. Por ejemplo, una bebé casi reción nacido que viaje en “Primera” tiene muchas probabilidades de sobrevivir:

```
female_baby <- data.frame(Sex="female", Age=0.1, SibSp=0, Pclass_category="Primera")
female_baby_prediction <- predict(logistic_model_optimized, female_baby, type = "resp
onse")
female_baby_prediction
```

```
##         1
## 0.9826912
```

En cambio un hombre adulto que viaje en “Tercera” no tiene casi posibilidades de sobrevivir:

```
adult_male <- data.frame(Sex="male", Age=30, SibSp=0, Pclass_category="Tercera")
adult_male_prediction <- predict(logistic_model_optimized, adult_male, type = "respon
se")
adult_male_prediction
```

```
##          1
## 0.09516064
```

Si este mismo hombre viajase en “Primera” tendría una probabilidad mucho más grande de sobrevivir:

```
adult_male_primera <- data.frame(Sex="male", Age=30, SibSp=0, Pclass_category="Primer
a")
adult_male_primera_prediction <- predict(logistic_model_optimized, adult_male_primer
a, type = "response")
adult_male_primera_prediction
```

```
##         1
## 0.5259651
```

Y si fuese una mujer que viajase en “Primera” clase tendría casi asegurada la supervivencia:

```
adult_female_primera <- data.frame(Sex="female", Age=30, SibSp=0, Pclass_category="Pr
imera")
adult_female_primera_prediction <- predict(logistic_model_optimized, adult_female_pri
mera, type = "response" )
adult_female_primera_prediction
```

```
##         1
## 0.9436214
```

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

Vamos a representar gráficamente algunas de las conclusiones que hemos obtenido.

```
library(ggplot2)
```

Como hemos visto, las proporciones relativas de supervivientes son muy diferentes en función de algunas de las variables, especialmente “Sex”, “Pclass” y “Embarked”. Esto se puede ver fácilmente usando tablas de frecuencia relativa de supervivencia (variable “Survived”) con respecto a estas variables. Podemos observar:

- Un 74% de las mujeres sobrevivieron, respecto a menos de un 19% de hombres.

```
freq_sex <- table(titanic_df$Survived, titanic_df$Sex)
prop.table(freq_sex, 2)
```

```
##
##        female      male
##   0 0.2596154 0.8110919
##   1 0.7403846 0.1889081
```

- Un 62.6% de los viajeros en “Primera” sobrevivieron, respecto a un 47.3% en “Segunda” y un 24.2% en “Tercera”.

```
freq_pclass <- table(titanic_df$Survived, titanic_df$Pclass)
prop.table(freq_pclass, 2)
```

```
##
##             1         2         3
##   0 0.3738318 0.5271739 0.7576375
##   1 0.6261682 0.4728261 0.2423625
```

- Un 55% de los viajeros que embarcaron en Cherbourg sobrevivieron, respecto a un 38.97% de los que embarcaron en Queenstown o un 33.7% de los que embarcaron en Southampton. Aunque como ya vimos esto se debía a que la mayoría de los pasajeros que embarcaron en Cherbourg viajaban en “Primera”.

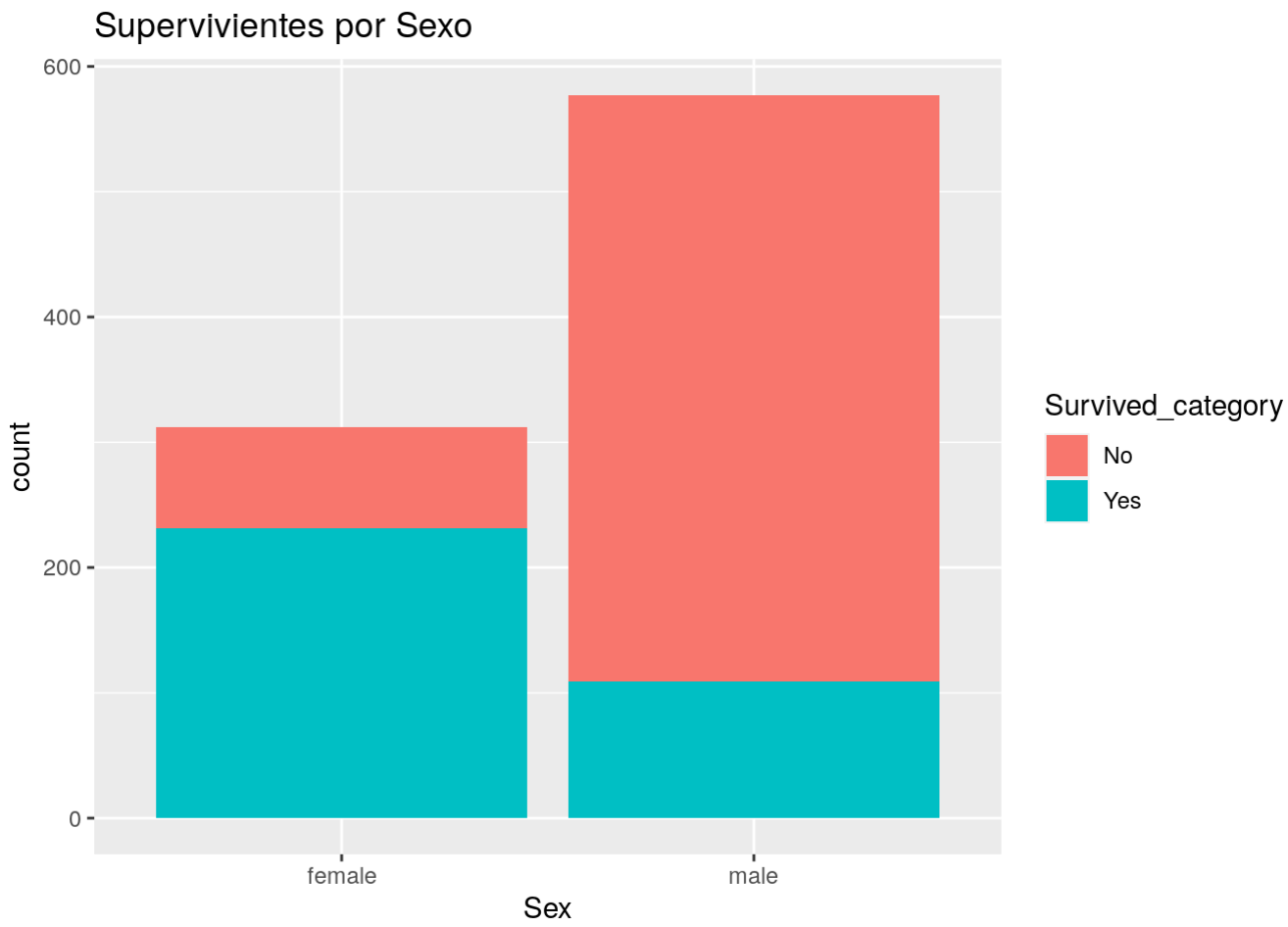
```
freq_embarked <- table(titanic_df$Survived, titanic_df$Embarked)
prop.table(freq_embarked, 2)
```

```
##
##             C         Q         S
##   0 0.4464286 0.6103896 0.6630435
##   1 0.5535714 0.3896104 0.3369565
```

Sin embargo, cuando miramos la distribución absoluta de los supervivientes obtenemos información también muy útil como:

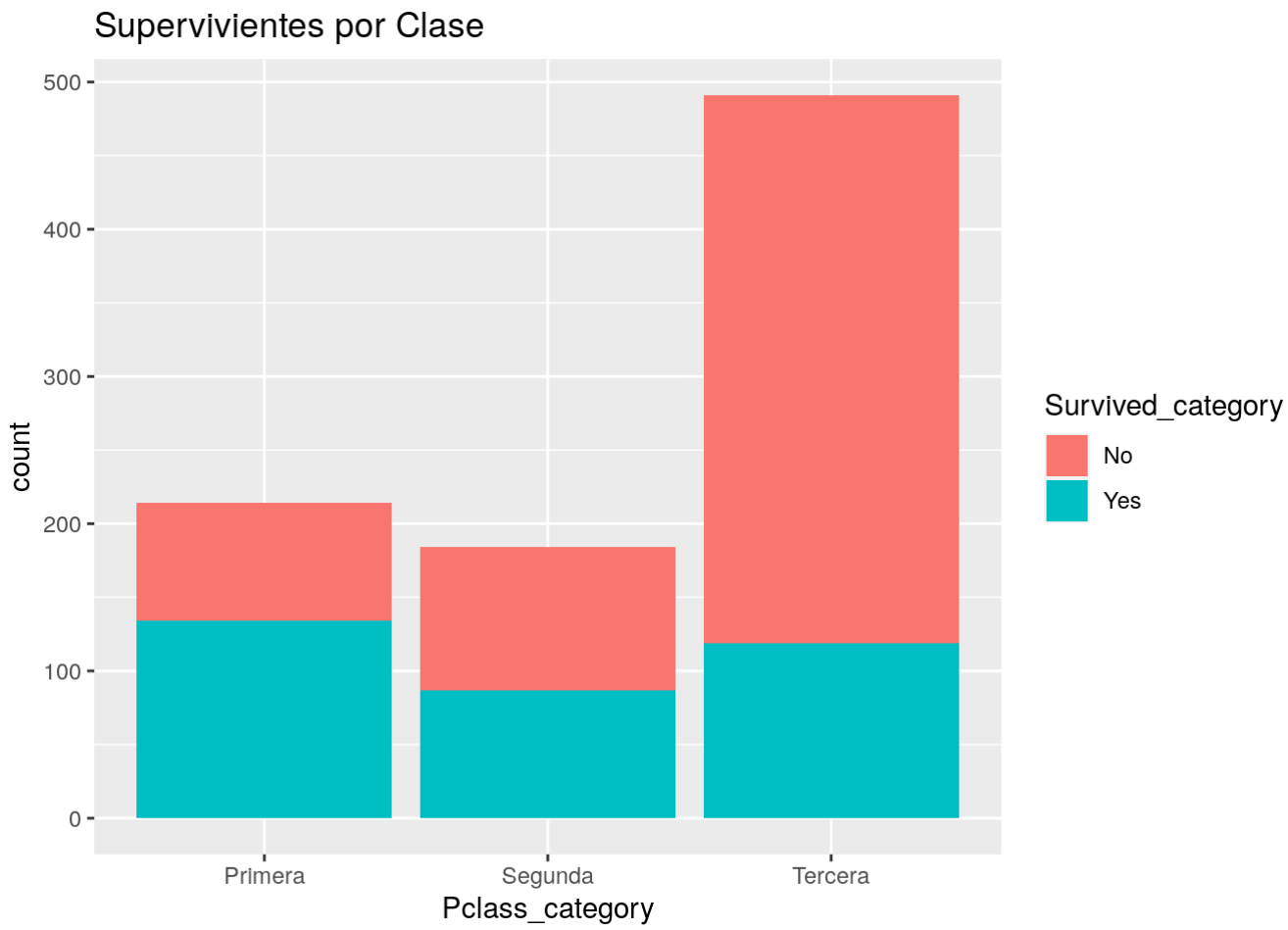
- La proporción de mujeres viajando en el Titanic era mucho menor que la de hombres. Sin embargo, sobrevivieron aproximadamente el doble de mujeres que de hombres.

```
ggplot(titanic_df, aes(Sex, fill=Survived_category)) + geom_bar() + ggtitle("Superviv
ientes por Sexo")
```



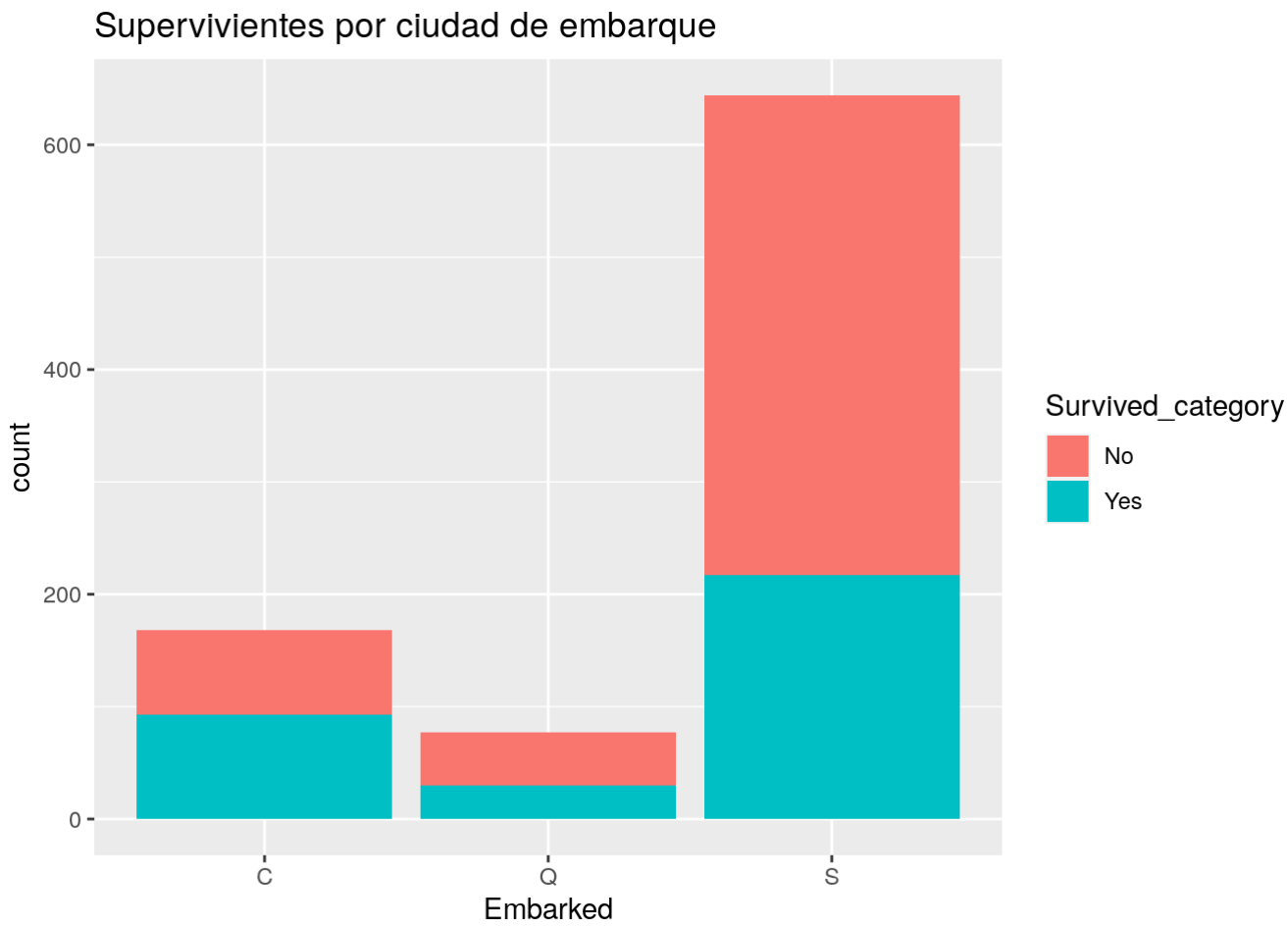

- Las probabilidades de sobrevivir viajando en “Tercera” como hemos comentado eran mucho más bajas que viajando en “Primera”. Sin embargo, en términos absolutos, la cantidad de supervivientes en cada clase es bastante similar.

```
ggplot(titanic_df, aes(Pclass_category, fill=Survived_category)) + geom_bar() + ggtit
le("Supervivientes por Clase")
```



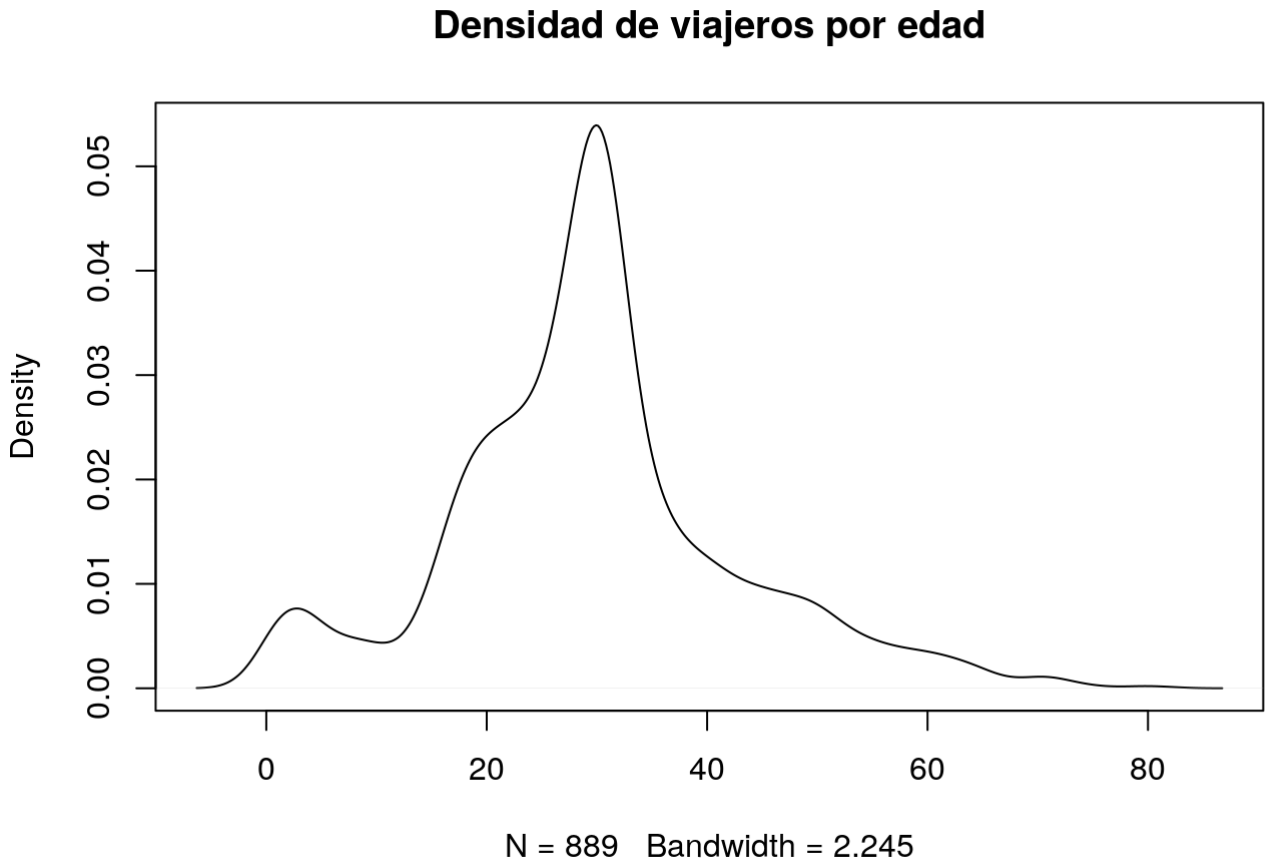

- En cuanto a la ciudad de embarque: aunque los viajeros que embarcaron en Chisbourg tenían más probabilidades de sobrevivir, en términos absolutos sobrevivieron el doble de viajeros de Southampton.

```
ggplot(titanic_df, aes(Embarked, fill=Survived_category)) + geom_bar() + ggtitle("Sup
ervivientes por ciudad de embarque")
```

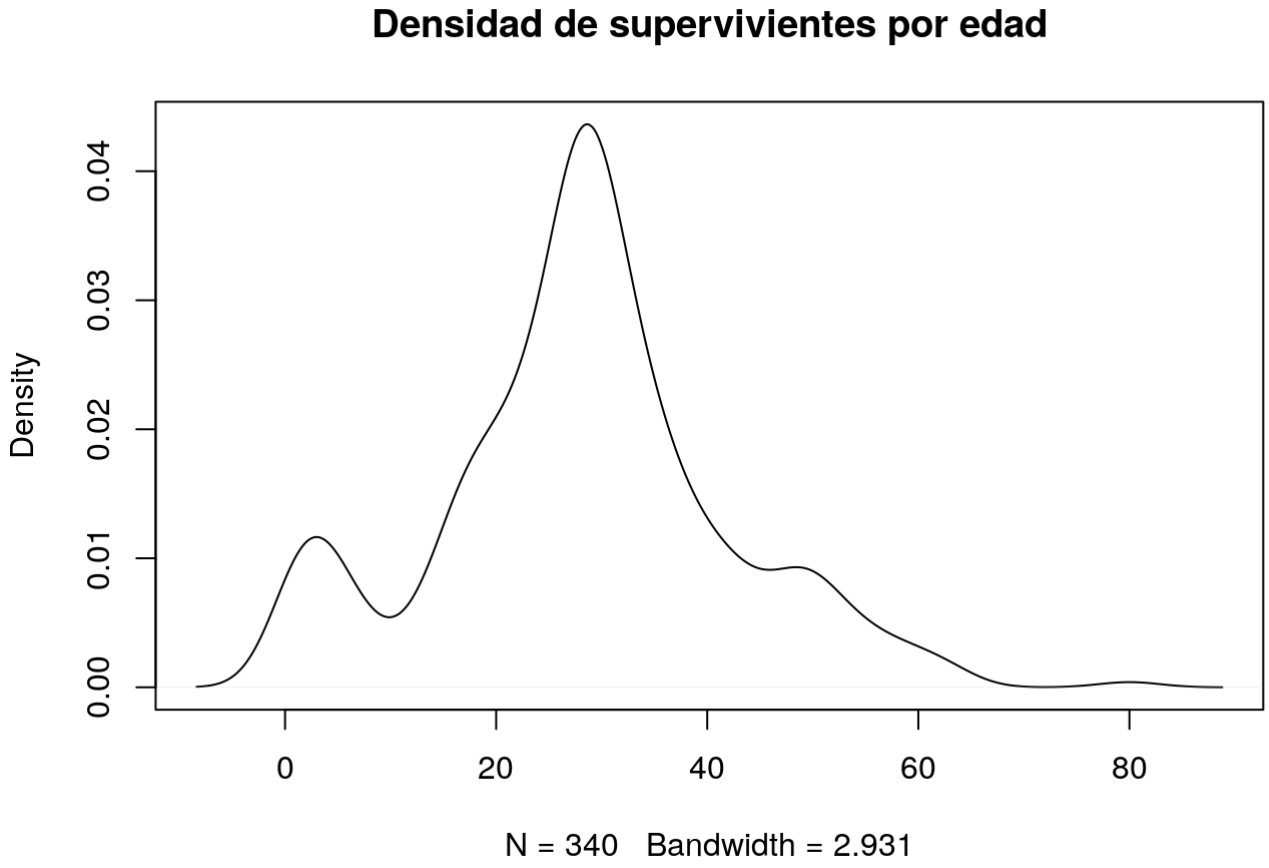



- Si comparamos la gráfica de densidad por edades de todos los viajeros y de los supervivientes, veremos que son bastante similares; aunque es perceptible que los niños pequeños tuvieron una probabilidad más alta de sobrevivir.

```
plot(density(titanic_df$Age), main="Densidad de viajeros por edad")
```




```
plot(density(titanic_df$Age[titanic_df$Survived==1]), main="Densidad de supervivientes por edad")
```



Por último, vamos a volver a nuestros modelo de regresión logística y vamos a evaluar su bondad usando las curbas ROC:

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
titanic_df$prob_survived = predict(logistic_model_optimized, titanic_df,
                              type="response")
g1=roc(titanic_df$Survived, titanic_df$prob_survived, data=titanic_df)
```

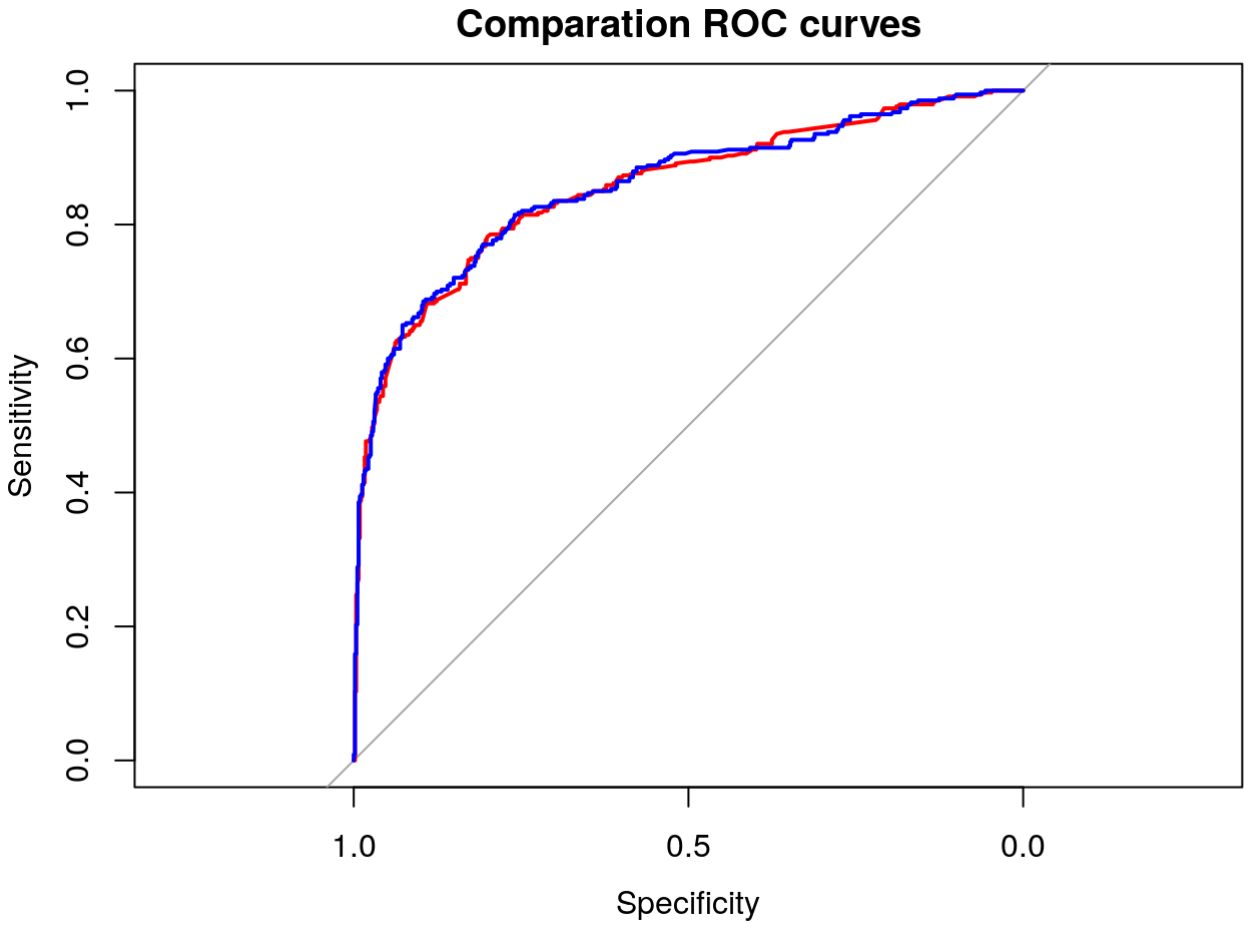
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
titanic_df$prob_survived = predict(logistic_model_all, titanic_df,
                              type="response")
g2=roc(titanic_df$Survived, titanic_df$prob_survived, data=titanic_df)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
plot(g1, main= "Comparation ROC curves", col="red")
lines(g2, col="blue")
```




Vemos que las curbas ROC de los 2 modelos son muy similares. En los 2 casos, las curbas se alejan de la diagonal, que representa la elección aleatoria, con lo que tendríamos modelos con una capacidad predictiva bastante buena.

# 6. Resolución del problema.

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

Los resultados obtenidos permiten contestar a las preguntas planteadas. Las conclusiones serían:

- Los contraste de hipótesis sobre las proporción nos han confirmado que hay una diferencia significativa a favor de las mujeres; por tanto, estas tuvieron más probabilidades de sobrevivir. Hemos comprobado también que aunque viajaban muchas menos mujeres (casi la mitad que hombres), sobrevivieron el doble de mujeres que de hombres. Un 74% de las mujeres sobrevivieron, respecto a menos de un 19% de los hombres a bordo.
- Este tipo de contraste también nos ha permitido comprobar que había una diferencia significativa en la proporción de supervivencia de los viajeros de “primera” respecto al resto de clases. La diferencia es importante respecto a “segunda” y mucho más respecto a “tercera”: un 62.6% de los viajeros en “Primera” sobrevivieron, respecto a un 47.3% en “Segunda” y un 24.2% en “Tercera”.
- En cuanto a la ciudad de embarque, hemos comprobado que los viajeros que embarcaron en Cherbourg tuvieron más probabilidades de sobrevivir que los viejeros que embarcaron en las otras ciudades. Sin embargo, cuando he ahondado en las características de estos viajeros, he comprobado que más del 50% de los viajeros que embarcaron en Cherbourg lo hicieron en “Primera” clase, cuando en Queenstown un 93.5% embarcaron en “Tercera” clase, y en Southampton menos de un 20% embarcaron en “Primera”. Esta sería la razón principal y no la ciudad de embarque en sí.
- También hemos visto en el estudio de las correlaciones y al crear los modelos de regresión logística que la variable “Age” correlaciona negativamente con “Survived”. Es decir, niños y jóvenes tuvieron más probabilidades de sobrevivir.

- Así mismo, la variable “SibSp” (número de hermanos y esposos) correlaciona positivamente con “Survived”, con lo que parece que los núcleos familiares amplios tuvieron más probabilidades de sobrevivir.
- Para finalizar, hemos creado varios modelos de regresión logística, que nos han confirmado, que como ya habíamos visto, las variables que más influencia tenían en el resultado eran la clase del billete, el sexo del viajero, la edad y la variable “SibSp”. Las curbas ROC de estos modelos nos muestran que tienen una capacidad de predicción bastante buena.

# Contribuciones

```
tabla_contribuciones <- data.frame(
  Contribuciones = c(
    "Investigación Previa",
    "Redacción de las respuestas",
    "Desarrollo código"
  ),
  Firma = c(
    "JF", "JF", "JF"
  )
)

kable(tabla_contribuciones)
```

| Contribuciones | Firma |
|---|---|
| Investigación Previa | JF |
| Redacción de las respuestas | JF |
| Desarrollo código | JF |