Análisis y limpieza de datos

Fausto De La Torre

enero 2021

Contents

Descripción del dataset

Trabajaremos con el conjunto de datos "Titanic" que recoge datos sobre la famosa embarcación. El set de datos sobre el que trabajaremos fue recogido del sitio de kaggle (https://www.kaggle.com/c/titanic/data); trabajeremos específicamente sobre los 891 registros del set de entrenamiento (train.csv) que contiene 12 variables descritas a continuación:

- **PassengerId** Identificador del pasajero
- **Survived** Indica si el pasajero sobrevivió (1) o no (2)
- **Pclass** Indica la clase del pasajero

- **Name** Nombre del pasajero
- **Sex** Sexo del pasajero
- **Age** Edad en años del pasajero
- **SibSp** Número de hermanos o cónyuges a bordo
- **Parch** Número de padres o hijos a bordo
- **Ticket** Número de ticket
- **Fare** Precio del ticket
- **Cabin** Número de cabina
- **Embarked** Puerto de embarcación: C = Cherbourg, Q = Queenstown, S = Southampton

Integración y selección de datos a analizar

En este conjunto de datos se busca analizar el impacto de las diferentes variables para determinar si el pasejero sobrevió o no al hundimiento del titanic. Lo que buscamos es conocer cuales son las variables que dictaminan la sobreviviencia de los pasajeros o si es una cuestión al azar.

Limpieza de datos

Primero procedemos a leer el set de datos y ha obtener estadísticas muy básicas

```
pasajeros <- read.csv("../data/titanic_train.csv", header=T, sep=",", stringsAsFactors = FALSE)
head(pasajeros)
```

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

```r
summary(pasajeros)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  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
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Valores vacíos

Se verifican los valores que se encuentran vacíos.

```r
print(colSums(pasajeros==""))
```

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

Se asigna "S" como valor para la variable "Embarked" en los dos registros que están vacíos ya que es el valor que está presente en la mayoría de registros.

```r
# Imputación de valores Embarked basados en la mayoría
pasajeros$Embarked[pasajeros$Embarked == ""] = "S"
```

Se elimina la variable "Cabin" ya que existen 687 registros vacíos que no aportarán demasiado en el análisis.

```r
pasajeros$Cabin = NULL
```

Se verifican los valores que se encuentran como NA.

```r
colSums(is.na(pasajeros))
```

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

Se calcula el valor medio para la edad tanto de hombres como de mujeres para asignar a los valores desconocidos.

```r
# # Imputación de valores Age basados en la media por sexo
pasajeros$Age[is.na(pasajeros$Age) & pasajeros$Sex == "male"] = mean(pasajeros$Age[!is.na(pasajeros$Age
pasajeros$Age[is.na(pasajeros$Age) & pasajeros$Sex == "female"] = mean(pasajeros$Age[!is.na(pasajeros$A
```

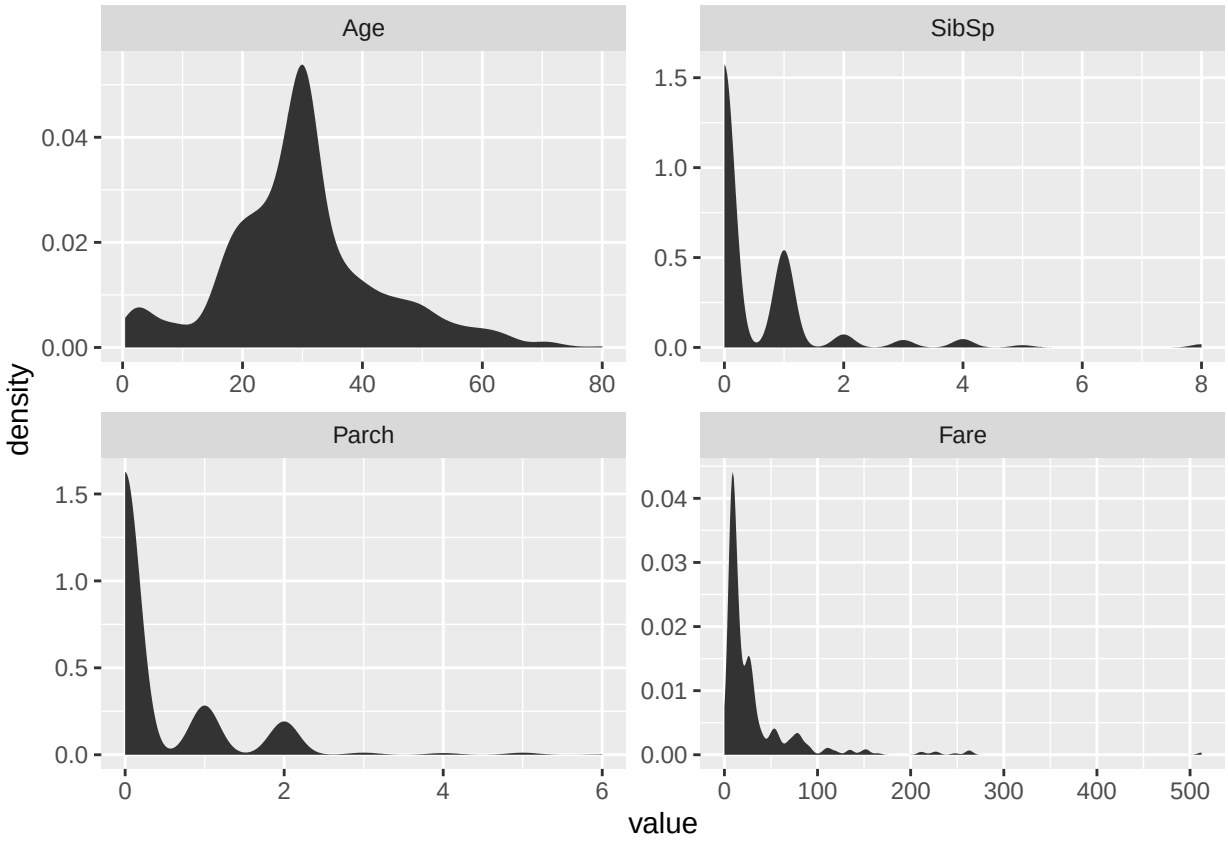
Valores Extremos

Primero vamos a ver las distribuciones de los valores continuos para ver posibles outliers y veremos también las la distribución de las variables discretas

```r
# Se hace un melt de las variables continuas para poder graficarlas
columns.numeric = c("Age", "SibSp", "Parch", "Fare")
melt.pasajeros <- melt(pasajeros[,columns.numeric])
```

```
## No id variables; using all as measure variables
```

```r
#Visualización de distribuciones de variables continuas
ggplot(melt.pasajeros) + aes(value) + stat_density() + facet_wrap(~variable, scales = "free")
```



Si vemos las distribuciones, podemos observar que existen valores extremos para las 4 variables continuas, sin embargo son valores que parecen ser válidos, no son errores.

Se elimina los registros del valor extremo de Fare cerca de 500

```
pasajeros = pasajeros[pasajeros$Fare < 300,]
```

Agregación de valores

De acuerdo al nombre de los pasajeros vamos a determinar el título que tienen (Mr., Mrs., etc.)

```
pasajeros$Title = str_extract(pasajeros$Name, "(?<=, ).+(?=\\. )")

# Se corrige el único registro fallido
pasajeros$Title[grepl("Mrs+\\.+ Martin +\\(+Elizabeth L", pasajeros$Title)] = "Mrs"

#Se agrupan los títulos
pasajeros$Title[pasajeros$Title == "Capt"] = "Crew"
pasajeros$Title[pasajeros$Title == "Col"] = "Crew"
pasajeros$Title[pasajeros$Title == "Major"] = "Crew"
pasajeros$Title[pasajeros$Title == "Rev"] = "Crew"
pasajeros$Title[pasajeros$Title == "Dr"] = "Crew"

pasajeros$Title[pasajeros$Title == "Master"] = "Royalty"
pasajeros$Title[pasajeros$Title == "Sir"] = "Royalty"
```

```
pasajeros$Title[pasajeros$Title == "Don"] = "Royalty"
pasajeros$Title[pasajeros$Title == "Lady"] = "Royalty"
pasajeros$Title[pasajeros$Title == "the Countess"] = "Royalty"
pasajeros$Title[pasajeros$Title == "Jonkheer"] = "Royalty"

pasajeros$Title[pasajeros$Title == "Ms"] = "Mrs"
pasajeros$Title[pasajeros$Title == "Mme"] = "Mrs"
pasajeros$Title[pasajeros$Title == "Mlle"] = "Miss"

unique(pasajeros$Title)
```

```
## [1] "Mr"      "Mrs"     "Miss"    "Royalty" "Crew"
```

Generación de un nuevo archivo con datos limpios

Se crea un nuevo archivo con los datos limpios

```
write.csv(pasajeros, '../data/out/titanic.csv')
```

Discretización de variables

Vamos a discretizar las variables continuas SibSp y Parch ya que contienen valores extremos a la derecha a los que vamos a agrupar en un solo grupo.

```
pasajeros$SibSp[pasajeros$SibSp >= 2] = '2+'
pasajeros$Parch[pasajeros$Parch >= 2] = '2+'
```

Discretizamos cuando tiene sentido y en función de cada variable.

```
apply(pasajeros,2, function(x) length(unique(x)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##         888           2           3         888           2          90
##       SibSp       Parch      Ticket        Fare    Embarked       Title
##           3           3         680         247           3           5
```

```
# Discretizamos las variables con pocas clases
cols<-c("Survived","Pclass","Sex","Embarked", "SibSp", "Parch", "Title")
for (i in cols){
  pasajeros[,i] <- as.factor(pasajeros[,i])
}
```

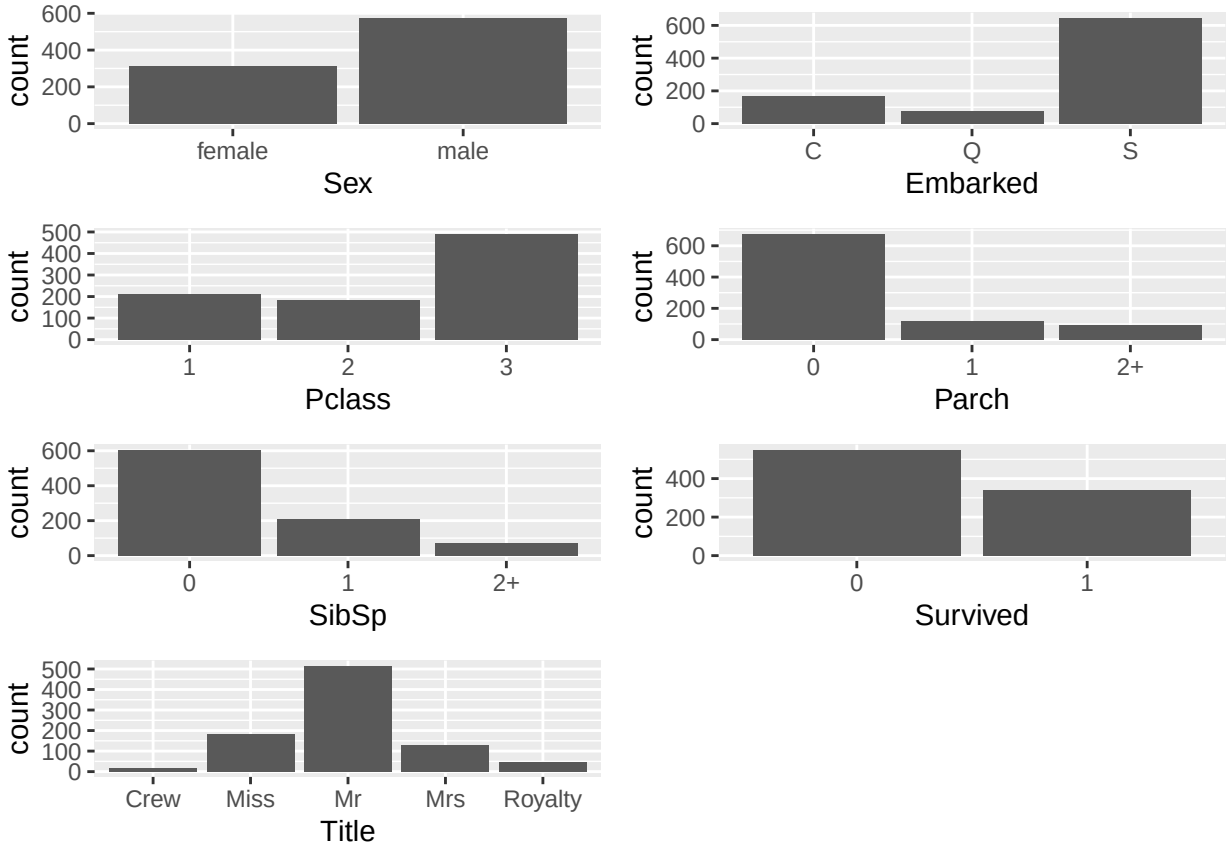
Después de los cambios, analizamos la nueva estructura del conjunto de pasajeros

```
str(pasajeros)
```

```
## 'data.frame':    888 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : Factor w/ 3 levels "0","1","2+": 2 2 1 2 1 1 1 3 1 2 ...
##  $ Parch      : Factor w/ 3 levels "0","1","2+": 1 1 1 1 1 1 1 2 3 1 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ Title      : Factor w/ 5 levels "Crew","Miss",..: 3 4 2 4 3 3 3 5 4 4 ...
```

Vemos la distribución de las variables discretas mediante un gráfico de barras

```
#Visualización de distribuciones de variables discretas
sex_plot = ggplot(pasajeros) + aes(Sex) + geom_bar()
embarked_plot = ggplot(pasajeros) + aes(Embarked) + geom_bar()
class_plot = ggplot(pasajeros) + aes(Pclass) + geom_bar()
parch_plot = ggplot(pasajeros) + aes(Parch) + geom_bar()
sibsp_plot = ggplot(pasajeros) + aes(SibSp) + geom_bar()
survived_plot = ggplot(pasajeros) + aes(Survived) + geom_bar()
title_plot = ggplot(pasajeros) + aes(Title) + geom_bar()

grid.arrange(sex_plot, embarked_plot, class_plot, parch_plot, sibsp_plot, survived_plot, title_plot, nc
```

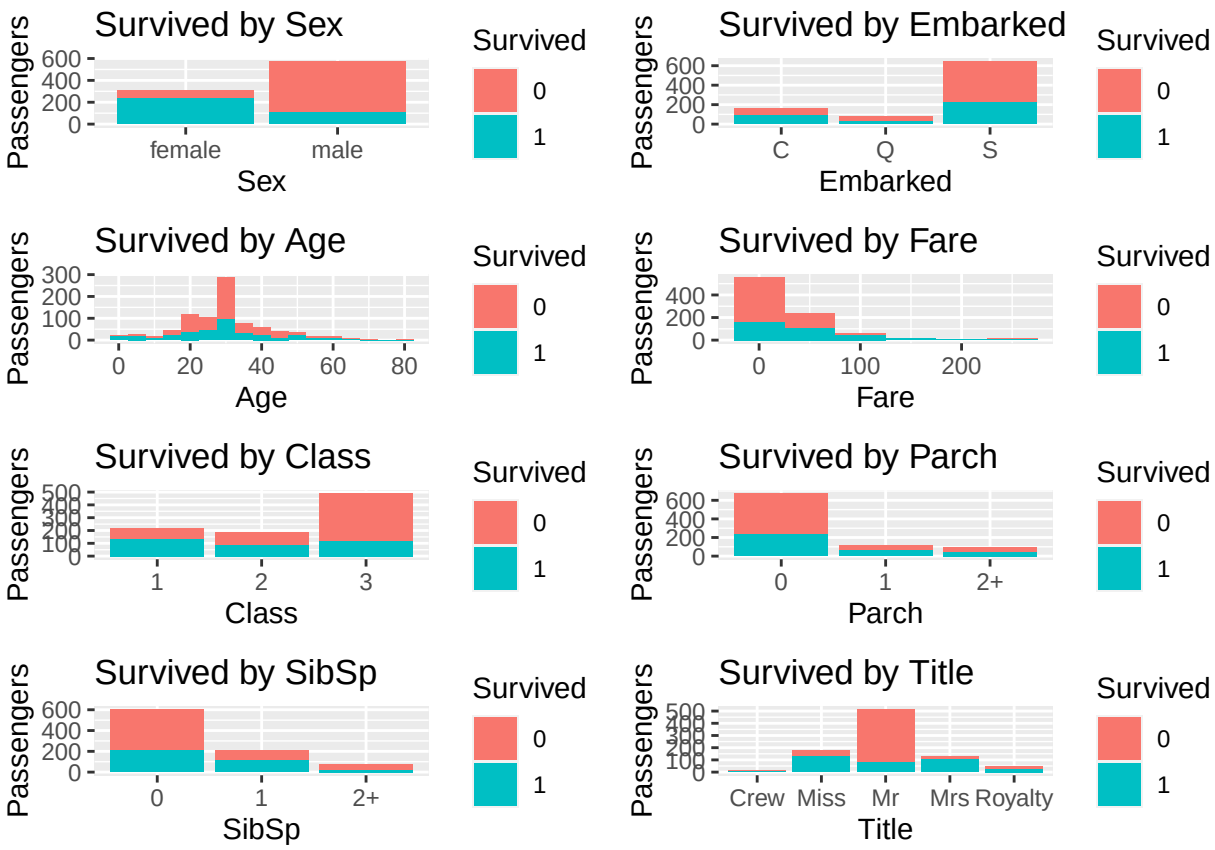


Análisis de datos

Selección de los grupos de datos a analizar

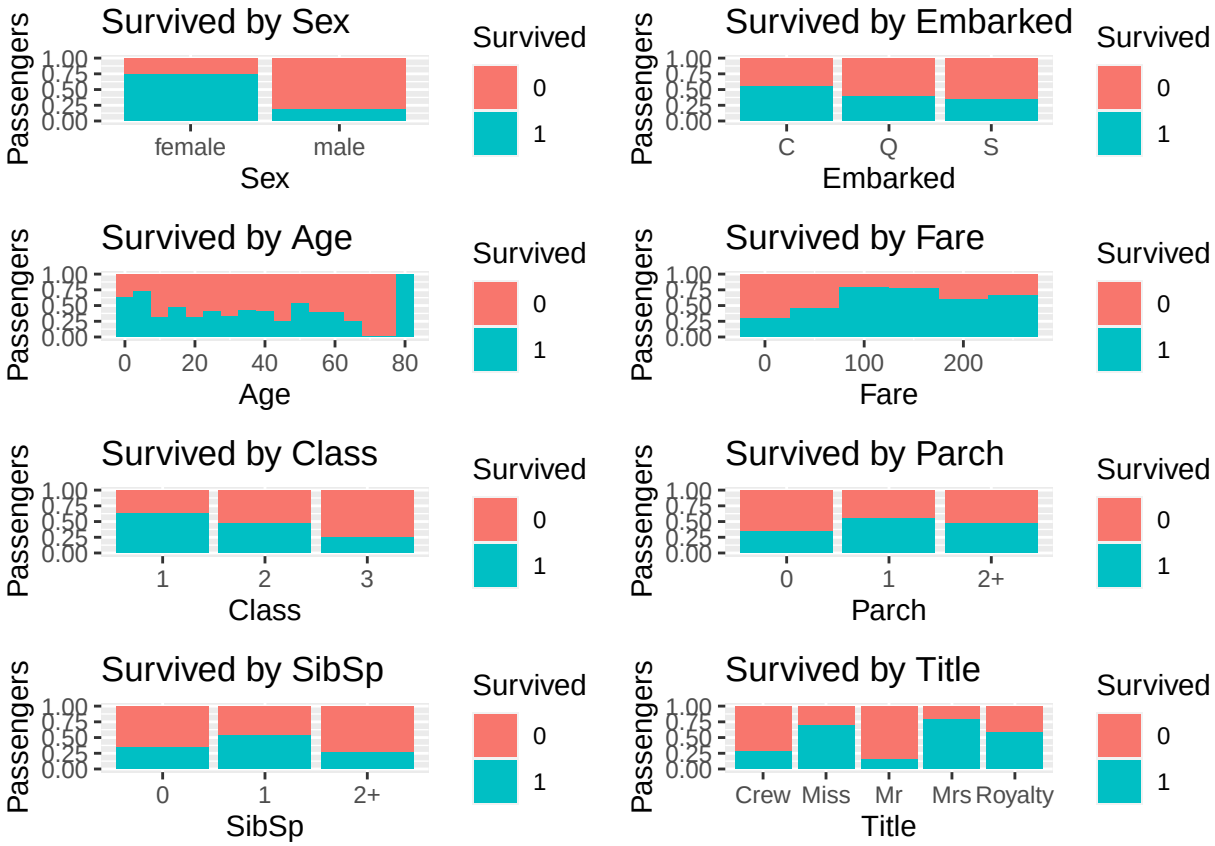
Vamos a realizar un análisis visual de los diferentes grupos en función de la variable "Survived"

```
sex_survived_plot = ggplot(data=pasajeros,aes(x=Sex,fill=Survived))+geom_bar() + labs(x="Sex", y="Passer

embarked_survived_plot = ggplot(data=pasajeros,aes(x=Embarked,fill=Survived))+geom_bar() + labs(x="Emba

age_survived_plot = ggplot(data=pasajeros,aes(x=Age,fill=Survived))+geom_histogram(binwidth = 5) + labs

fare_survived_plot = ggplot(data=pasajeros,aes(x=Fare,fill=Survived))+geom_histogram(binwidth = 50) + la

class_survived_plot = ggplot(data=pasajeros,aes(x=Pclass,fill=Survived)) + geom_bar() + labs(x="Class",

parch_survived_plot = ggplot(data=pasajeros,aes(x=Parch,fill=Survived)) + geom_bar() + labs(x="Parch", y

sibsp_survived_plot = ggplot(data=pasajeros,aes(x=SibSp,fill=Survived)) + geom_bar() + labs(x="SibSp", y

title_survived_plot = ggplot(data=pasajeros,aes(x=Title,fill=Survived)) + geom_bar() + labs(x="Title", y


grid.arrange(sex_survived_plot, embarked_survived_plot, age_survived_plot, fare_survived_plot, class_su
```



Ahora vemos los gráficos en terminos relativos.

```
sex_survived_plot = ggplot(data=pasajeros,aes(x=Sex,fill=Survived))+geom_bar(position="fill") + labs(x=

embarked_survived_plot = ggplot(data=pasajeros,aes(x=Embarked,fill=Survived))+geom_bar(position="fill")

age_survived_plot = ggplot(data=pasajeros,aes(x=Age,fill=Survived))+geom_histogram(binwidth = 5, positi

fare_survived_plot = ggplot(data=pasajeros,aes(x=Fare,fill=Survived))+geom_histogram(binwidth = 50, posi

class_survived_plot = ggplot(data=pasajeros,aes(x=Pclass,fill=Survived)) + geom_bar(position="fill") + l

parch_survived_plot = ggplot(data=pasajeros,aes(x=Parch,fill=Survived)) + geom_bar(position="fill") + la

sibsp_survived_plot = ggplot(data=pasajeros,aes(x=SibSp,fill=Survived)) + geom_bar(position="fill") + la

title_survived_plot = ggplot(data=pasajeros,aes(x=Title,fill=Survived)) + geom_bar(position="fill") + la

grid.arrange(sex_survived_plot, embarked_survived_plot, age_survived_plot, fare_survived_plot, class_su
```



Los grupos de interés que exploraremos son por sexo, puerto de embarque y clase

```
pasajeros.hombres = pasajeros[pasajeros$Sex == "male",]
pasajeros.mujeres = pasajeros[pasajeros$Sex == "female",]

pasajeros.clase1 = pasajeros[pasajeros$Pclass == 1,]
pasajeros.clase2 = pasajeros[pasajeros$Pclass == 2,]
pasajeros.clase3 = pasajeros[pasajeros$Pclass == 3,]
```

```
pasajeros.cherbourg = pasajeros[pasajeros$Embarked == "C",]
pasajeros.queenstown = pasajeros[pasajeros$Embarked == "Q",]
pasajeros.southampton = pasajeros[pasajeros$Embarked == "S",]
```

Comprobación de la normalidad y homogeneidad de la varianza

Podemos observar que la distribución de las dos variables continuas no son normales, sin embargo lo vamos a verificar mediante la prueba de Shapiro-Wilk en la que la hipótesis nula nos dice que la distribución es normal.

```
p_value_age = shapiro.test(pasajeros$Age)$p.value
p_value_fare = shapiro.test(pasajeros$Fare)$p.value

print(sprintf("p-value para Age del test de shapiro: %.6e", p_value_age))
```

```
## [1] "p-value para Age del test de shapiro: 3.789210e-14"
```

```
print(sprintf("p-value para Fare del test de shapiro: %.6e", p_value_fare))
```

```
## [1] "p-value para Fare del test de shapiro: 7.407402e-41"
```

Los valores de p menores al nivel de significancia de 5% nos dice que podemos descartar la hipótesis nula, por lo tanto concluimos que las distribuciones no son normales para Age y Fare.

Mediante la prueba de homoscedasticidad para determinar la homogeneidad o no de las varianzas, utilizaremos F-test cuya hipótesis nula nos dice que existe homoscedasticidad (varianzas similares) para la variable Fare

```
var.test(pasajeros.hombres$Fare, pasajeros.mujeres$Fare)
```

```
##
##  F test to compare two variances
##
## data:  pasajeros.hombres$Fare and pasajeros.mujeres$Fare
## F = 0.38896, num df = 574, denom df = 312, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3190577 0.4714808
## sample estimates:
## ratio of variances
##          0.3889617
```

El p-value del F-test es menor que el nivel de significancia 0.05, por lo que podemos concluir que se rechaza la hipótesis nula, que nos dice que hay una diferencia significativa en las varianzas de las dos muestras. Las diferencias ente

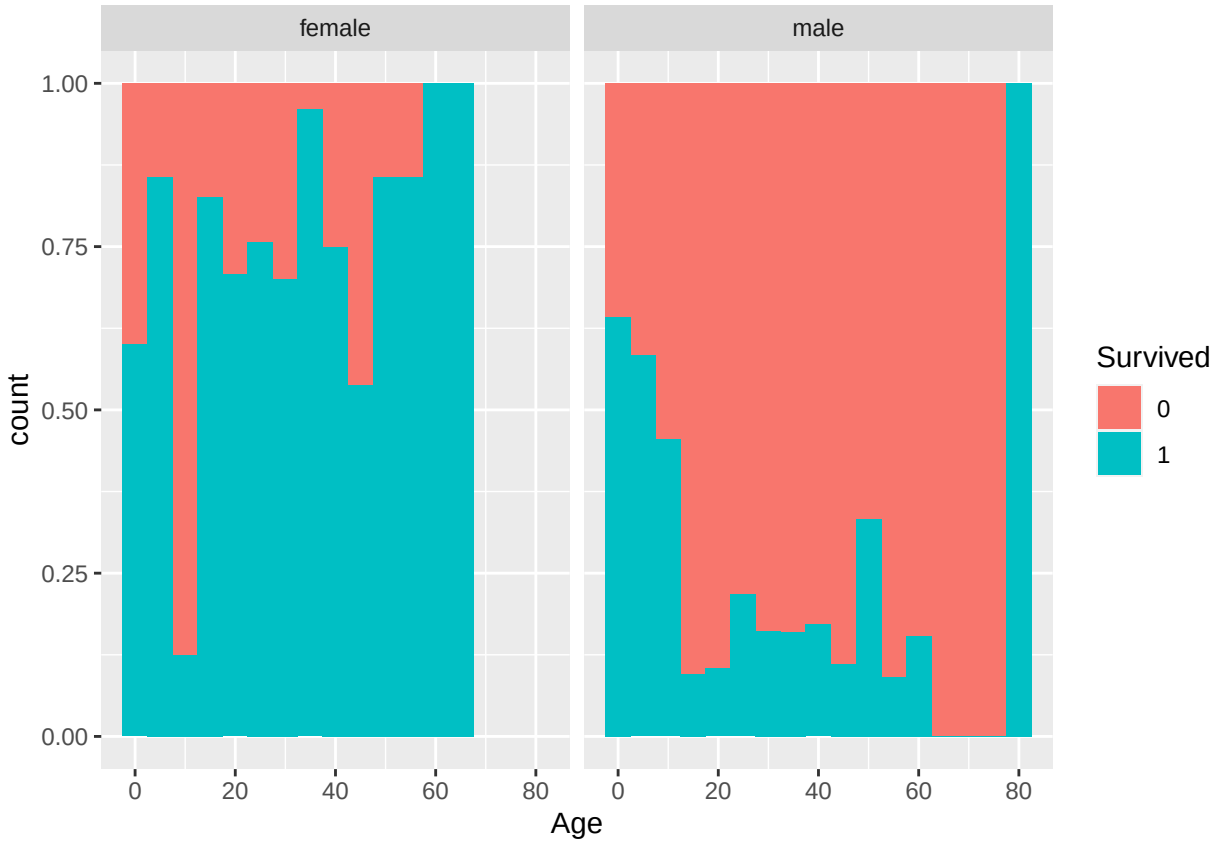
Pruebas estadísticas

Lo que buscamos es determinar que variables son las que más influyen en determinar si un pasajero sobrevivió o no al hunidmiento del titanic.

Primero vamos a identificar las relaciones entre algunas variables

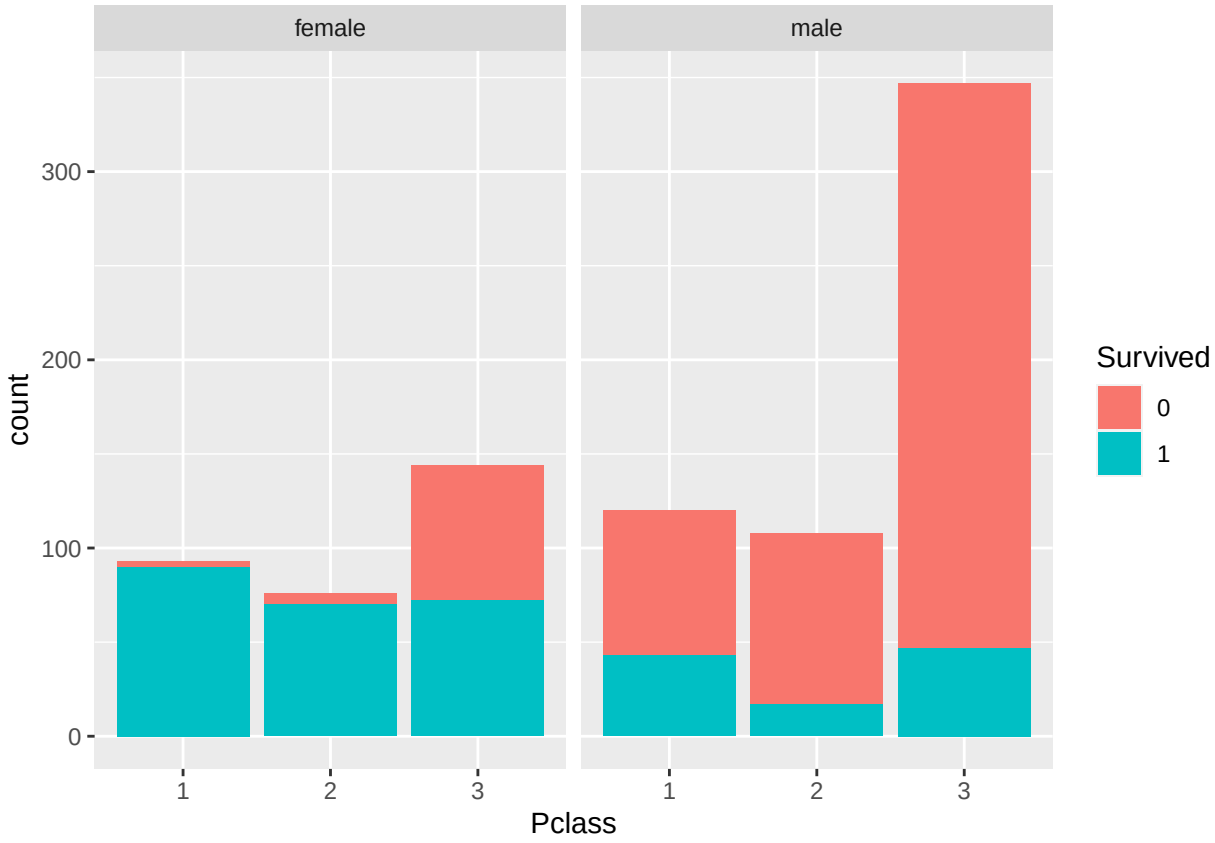
```
ggplot(data = pasajeros, aes(x=Age, fill=Survived)) + geom_histogram(binwidth = 5, position = "fill") +
```

```
## Warning: Removed 6 rows containing missing values (geom_bar).
```



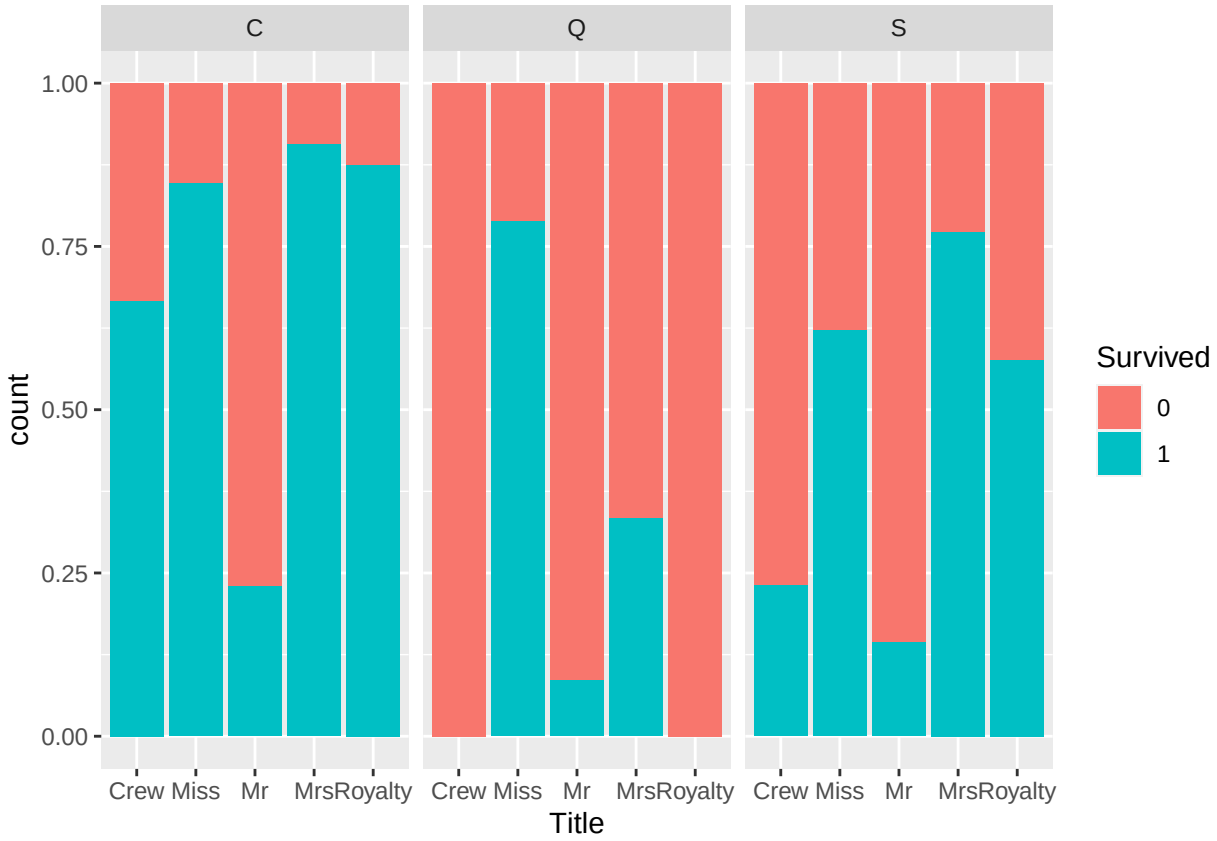
La mayor parte de sobrevivientes son mujeres o niños (hombres y mujeres)

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



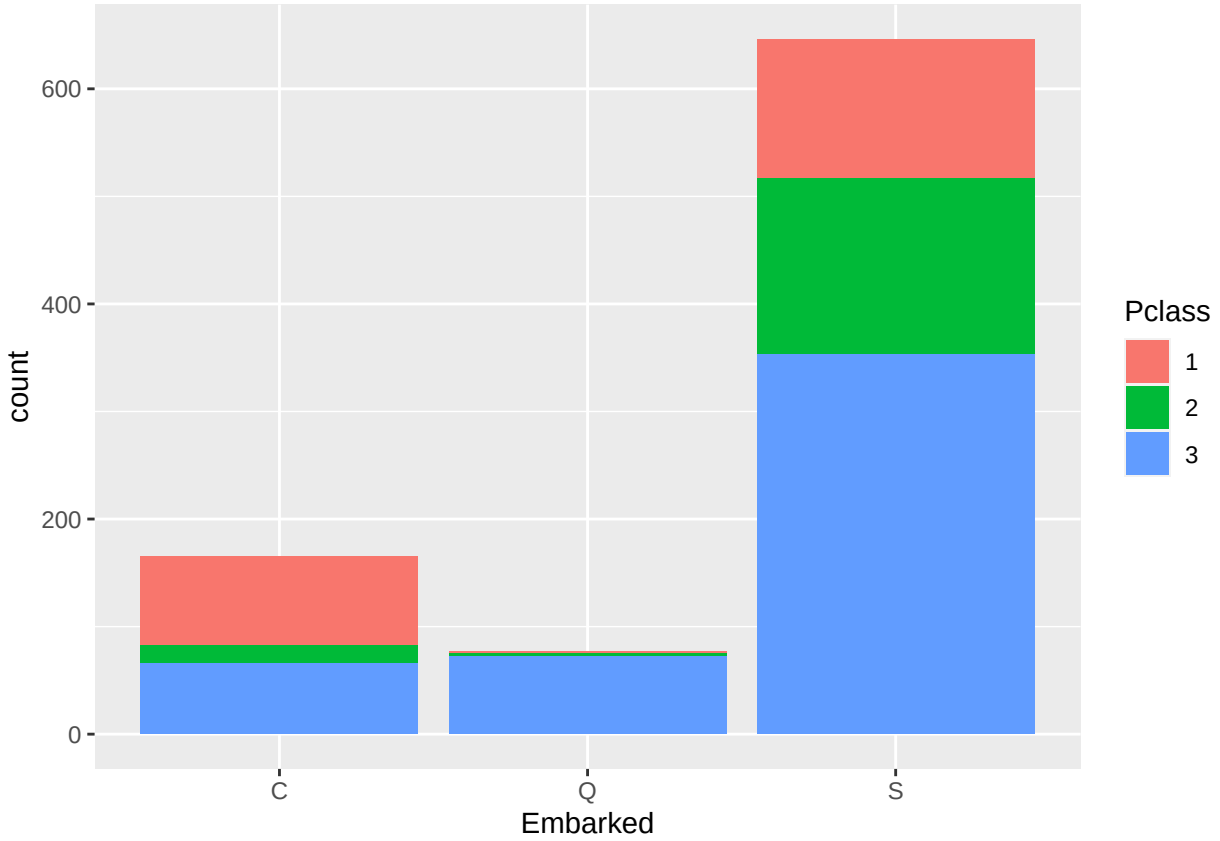
- Casi todas las mujeres de primera y segunda clase sobrevivieron, sin embargo el número de mujeres en 3ra clase es similar al de las otras clases.
- Los hombres en general se sacrificaron, sin embargo se evidencia que si los hombres estaban en 1ra clase tenían más probabilidades de sobrevivir.

```
ggplot(data = pasajeros, aes(x=Title, fill=Survived)) + geom_bar(position = "fill") + facet_wrap(~Embarl
```



- A excepción de los Señores (Mr.), las personas que se embarcaron en Cherbourg tuvieron más probabilidades de sobrevivir, es así que la tripulación que se embarcó en dicho puerto es la que sobrevive casi en su mayoría.
- Un dato curioso, la realeza embarcada en Queenstown no sobrevivió.

```
ggplot(data = pasajeros, aes(x=Embarked, fill=Pclass)) + geom_bar()
```



Vamos a generar un árbol de decisión del cual sacaremos ciertas reglas utilizando las variables “Age”, “Embarked”, “Pclass”, “Sex”

```
X = pasajeros[,c("Age", "Embarked", "Pclass", "Sex")]
y = pasajeros$Survived
model <- C50::C5.0(X, y,rules=TRUE )
summary(model)
```

```
##
## Call:
## C5.0.default(x = X, y = y, rules = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Mon Jan  4 12:04:49 2021
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 888 cases (5 attributes) from undefined.data
##
## Rules:
##
## Rule 1: (538/86, lift 1.4)
##  Age > 13
##  Sex = male
##  ->  class 0  [0.839]
```

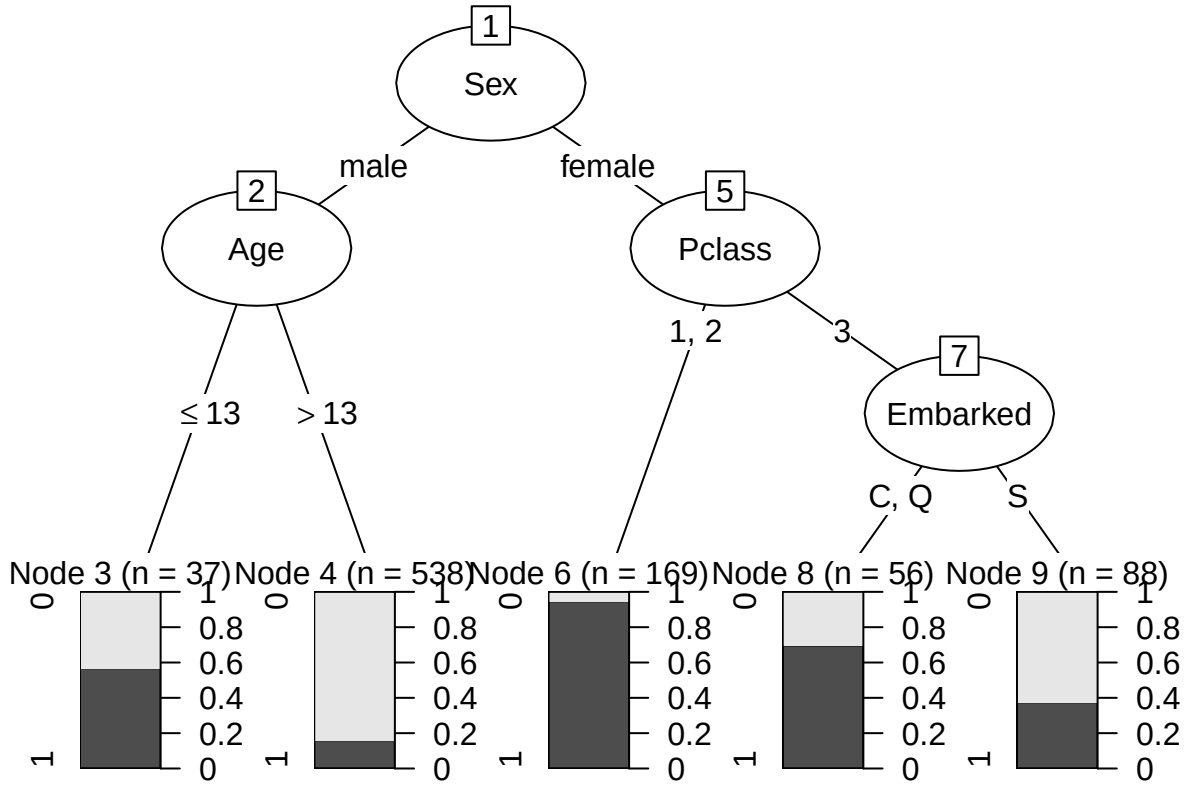
```
##
## Rule 2: (491/119, lift 1.2)
##  Pclass = 3
##  ->  class 0  [0.757]
##
## Rule 3: (169/9, lift 2.5)
##  Pclass in {1, 2}
##  Sex = female
##  ->  class 1  [0.942]
##
## Rule 4: (108/18, lift 2.2)
##  Embarked in {C, Q}
##  Sex = female
##  ->  class 1  [0.827]
##
## Rule 5: (71/29, lift 1.5)
##  Age <= 13
##  ->  class 1  [0.589]
##
## Default class: 0
##
##
## Evaluation on training data (888 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##       5  154(17.3%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     523    26    (a): class 0
##     128   211    (b): class 1
##
##
##  Attribute usage:
##
##   85.92% Sex
##   74.32% Pclass
##   68.58% Age
##   12.16% Embarked
##
##
## Time: 0.0 secs
```

Se puede ver que la variable más determinante es Sex, seguido por la clase y por la edad.

- Si vemos la regla 3 se puede ver que casi todas las mujeres en 1ra y 2da clase sobrevivieron.
- Si vemos la regla 1 se puede ver que casi todos los hombres adultos no sobrevicvieron al naufragio de la embarcación.

Vamos a visualizar el árbol para mayor claridad

```
model <- C50::C5.0(X, y)
plot(model)
```



Este modelo cometió 154 errores, lo que representa 17.3% de la muestra. Esto nos da suficiente explicabilidad para nuestro análisis.

A continuación crearemos un modelo similar pero escogiendo la variable Title en lugar de Sex eliminando solamente utilizaremos las variables "Embarked", "Pclass", "Title". Queremos ver si podemos tener más insights.

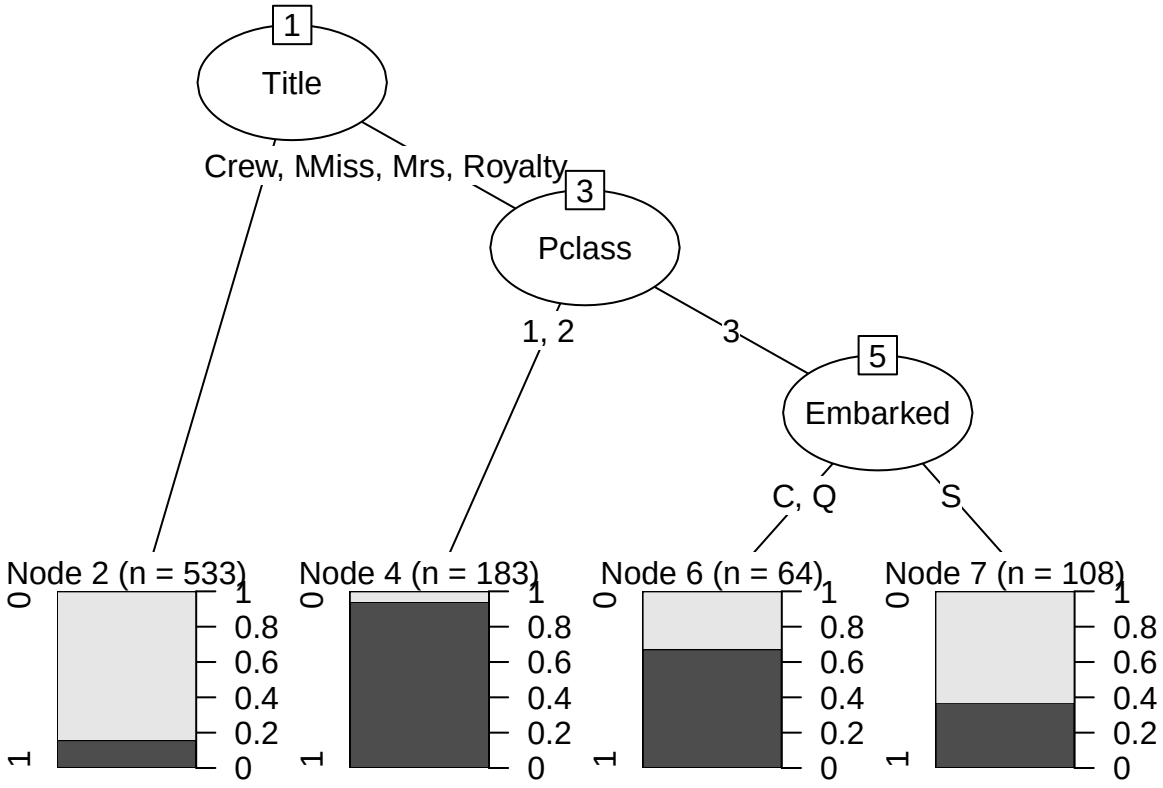
```
X = pasajeros[,c("Embarked", "Pclass", "Title")]
y = pasajeros$Survived
model <- C50::C5.0(X, y,rules=TRUE )
summary(model)
```

```
##
## Call:
## C5.0.default(x = X, y = y, rules = TRUE)
##
##
## C5.0 [Release 2.07 GPL Edition]      Mon Jan  4 12:04:49 2021
## -------------------------------
##
## Class specified by attribute 'outcome'
##
## Read 888 cases (4 attributes) from undefined.data
```

```
##
## Rules:
##
## Rule 1: (533/84, lift 1.4)
##  Title in {Crew, Mr}
##  ->  class 0  [0.841]
##
## Rule 2: (353/67, lift 1.3)
##  Embarked = S
##  Pclass = 3
##  ->  class 0  [0.808]
##
## Rule 3: (183/11, lift 2.4)
##  Pclass in {1, 2}
##  Title in {Miss, Mrs, Royalty}
##  ->  class 1  [0.935]
##
## Rule 4: (119/23, lift 2.1)
##  Embarked in {C, Q}
##  Title in {Miss, Mrs, Royalty}
##  ->  class 1  [0.802]
##
## Default class: 0
##
##
## Evaluation on training data (888 cases):
##
##          Rules
##    ----------------
##      No      Errors
##
##       4  156(17.6%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     517    32    (a): class 0
##     124   215    (b): class 1
##
##
##  Attribute usage:
##
##   87.84% Title
##   60.36% Pclass
##   53.15% Embarked
##
##
## Time: 0.0 secs
```

Este modelo es

```
model <- C50::C5.0(X, y)
plot(model)
```



Conclusiones

Luego de haber analizado los datos relacionados en función de la variable “survived” podemos obtener las siguientes conclusiones:

- Los hombres, especialmente los de 2da y tercera clase fueron los que más se sacrificaron
- La realeza, y las mujeres de 1ra y 2da clase fueron las pasajeras a salvar mayoritariamente así como los niños y niñas.
- La mayor parte de los pasajeros se embarcaron en Southampton, dichos pasajeros murieron en su mayoría, sin embargo en porcentaje es muy similar a los otros puertos, esto nos dice que no hubo una predilección por puerto para la sobreviviencia.
- Casi todas las personas que se embarcaron en Q fueron a 3ra clase.