PRA2 - Análisis de datos Titanic-Kaggle

Javier Pérez Córdova

9/6/2020

Contents

1. Descripción del dataset

El conjunto de datos está disponible a partir del enlace de kaggle referente a la competición del titanic. Para este análisis hemos decidido actuar con el dataset completo (train+test) dejando fuera del análisis inicial la variable relativa a la supervivencia, que será tenida en cuenta en el apartado de análisis. El dataset de test será preparado, pero solo usado en el caso de la regresión para probar nuestro modelo.
Primero realizamos unas exploración general de los datos que tenemos.

En total tenemos datos sobre 1309 pasajeros y 12 variables referentes a ellos que se explican a continuación:

- PassengerId: ID del pasajero.
- Survived: Si el pasajero ha sobrevivido o no al hundimiento (0=no, 1=yes).
- Pclass: Clase en la que viajaba en el titanic (1, 2, 3), siendo mayor clase a menor valor.
- Name: Nombre del pasajero.
- Sex: sexo del pasajero.
- Age: edad del pasajero en años.
- SibSp: número (cantidad) de relación familiar (Sibiling/ Spouse).
- Parch: número (cantidad) de relación familiar (Parent/child).
- Ticket: número del ticket.
- Fare: precio del ticket.
- Cabin: número de la cabina.
- Embarked: Puerto en el que ha embarcado (Cherbourg, Queenstown,Southampton).

De los 1309 pasajeros, tenemos el resultado de supervivencia para 891 de ellos.
El objetivo perseguido con este conjunto de datos es llegar a conocer que variables fueron las más influyentes a la hora de sobrevivir a la tragedia del titanic, por ejemplo, si el nivel socioeconómico el sexo o la edad tuvieron impacto en la salvación.

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

Lo primero será cargar los registros y combinarlos para llegar a las 1309 entradas, y hacer una observación inicial de los valores que tenemos.

```
#Cargamos los datos y combinamos para poder analizar los datos de entrada que tenemos
test_split=read.csv('../datasets/test.csv')
train_split=read.csv('../datasets/train.csv')
#Se crea columna vacia de Survived para la porción de test
test_split['Survived']<-NA
#Se combinan ambos datasets para realizar la posterior
#limpieza de datos y los dos primero apartados del Análisis
full_data= rbind(train_split, test_split)
#Resumen de la estructura de los datos
str(full_data, width = 80, strict.width = "cut")
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 109 191 358 27..
##  $ 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 ...
##  $ Ticket     : Factor w/ 929 levels "110152","110413",..: 524 597 670 50 473 ..
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 187 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 ..
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
summary(full_data)
```

```
##   PassengerId      Survived          Pclass     
##  Min.   :   1   Min.   :0.0000   Min.   :1.000  
##  1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000  
##  Median : 655   Median :0.0000   Median :3.000  
##  Mean   : 655   Mean   :0.3838   Mean   :2.295  
##  3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000  
##  Max.   :1309   Max.   :1.0000   Max.   :3.000  
##                 NA's   :418                     
##                                 Name          Sex           Age       
##  Connolly, Miss. Kate                :   2   female:466   Min.   : 0.17  
##  Kelly, Mr. James                    :   2   male  :843   1st Qu.:21.00  
##  Abbing, Mr. Anthony                 :   1                Median :28.00  
##  Abbott, Mr. Rossmore Edward         :   1                Mean   :29.88  
##  Abbott, Mrs. Stanton (Rosa Hunt):   1                3rd Qu.:39.00  
##  Abelson, Mr. Samuel                 :   1                Max.   :80.00  
##  (Other)                             :1301                NA's   :263    
##      SibSp            Parch           Ticket          Fare        
##  Min.   :0.0000   Min.   :0.000   CA. 2343:  11   Min.   :  0.000  
##  1st Qu.:0.0000   1st Qu.:0.000   1601    :   8   1st Qu.:  7.896  
##  Median :0.0000   Median :0.000   CA 2144 :   8   Median : 14.454  
##  Mean   :0.4989   Mean   :0.385   3101295 :   7   Mean   : 33.295  
##  3rd Qu.:1.0000   3rd Qu.:0.000   347077  :   7   3rd Qu.: 31.275  
##  Max.   :8.0000   Max.   :9.000   347082  :   7   Max.   :512.329  
##                                   (Other) :1261   NA's   :1
```

```
##              Cabin      Embarked
##                 :1014    :  2
##  C23 C25 C27    :   6   C:270
##  B57 B59 B63 B66:   5   Q:123
##  G6             :   5   S:914
##  B96 B98        :   4
##  C22 C26        :   4
##  (Other)        : 271
```

Como bien se ha comentado en el apartado anterior, se seleccionan los 1309 registro y las variables Pclass, sex, Age, SibSp,Parch, Fare,Embarked y Survived ya que se consideran, tras la primera exploración visual, como las variables que nos podrán aportar algo en nuestro análisis y en análisis derivado y accesorios. A continuación, se muestra el código para eliminar las variables carente de interés, mediante el índice de la columna a la que pertenecen.

```
#Se eliminan las columnas que no nos interesan
full_data<-full_data[ -c(1,4,9,11) ]
colnames(full_data)
```

```
## [1] "Survived" "Pclass"   "Sex"      "Age"      "SibSp"    "Parch"    "Fare"
## [8] "Embarked"
```

Para la mayor parte del análisis se usarán los 891 valores que tienen un valor Survived asociado, mientras que, para la parte final de regresión, se utilizarán los 418 con el ánimo de llegar a predecir su valor Survived asociado.

Vistos los tipos de datos, debemos transformar las variables Survived y Pclass, ya que los valores que contienen actúan como categorías y no como números.

```
#Se transforman en factor aquellas variables seleccionadas
full_data$Survived<-as.factor(full_data$Survived)
full_data$Pclass<-as.factor(full_data$Pclass)
str(full_data)
```

```
## 'data.frame':    1309 obs. of  8 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 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 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

3. Limpieza de los datos

3.1. Localización y tratamiento de valores nulos y vacíos

Una vez cargados todos los valores, procedemos a revisar que las columnas seleccionadas no tiene valores nulos o vacíos.

```
#Se revisa el dataset para ver la incidencia de valores nulos y NA,
#también podríamos hacerlo aplicndo summary(full_data)
colSums(is.na(full_data))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##      418        0        0      263        0        0        1        0
```

```
colSums((full_data==''))
```

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

Podemos observar como la variable Survived tiene valores NA, pero son los que hemos introducido nosotros manualmente para la unión del dataset, por lo que los mantendremos invariables.
En el caso de la variable Age vemos como 263 de los registros presentan el valor NA. Para solucionar esta problemática usaremos la función kNN() del paquete VIM, siguiendo las recomendaciones de los apuntes debido a su robusted y sencillez de uso. También observamos que nos falta un valor de Fare, por lo que aplicaremos la misma técnica.

```
suppressWarnings(suppressMessages(library(VIM)))
#Se realiza la imputación de valores NA
full_data$Age<-kNN(full_data)$Age
full_data$Fare<-kNN(full_data)$Fare
colSums(is.na(full_data))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##      418        0        0        0        0        0        0        0
```

Para el caso de valores vacíos en la variable embarked, se ha decidido asignar a NA y realizar la imputación con KNN.

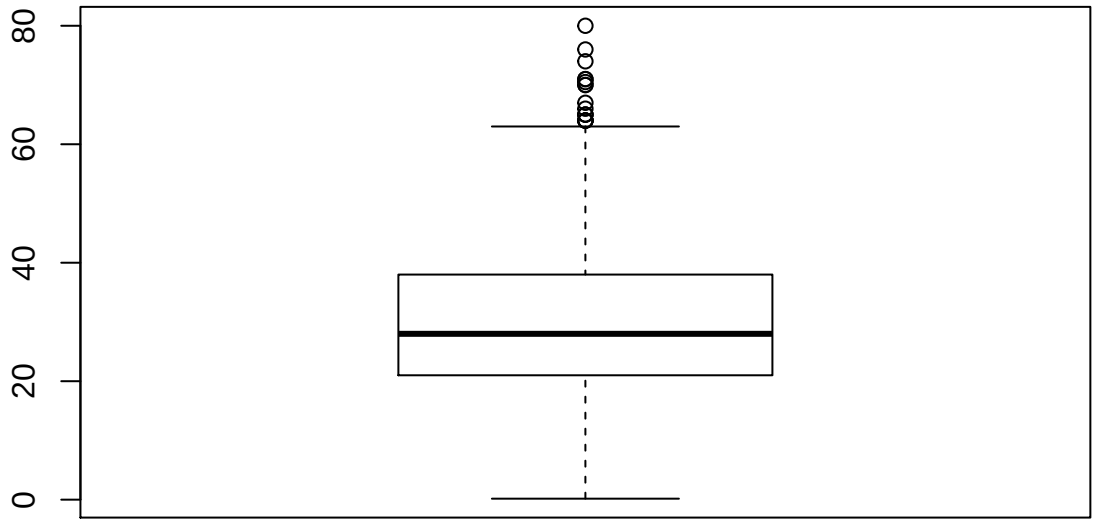
```
full_data$Embarked[full_data$Embarked==""]=NA
full_data$Embarked<-kNN(full_data)$Embarked
colSums((full_data==''))
```

```
## Survived   Pclass      Sex      Age    SibSp    Parch     Fare Embarked
##       NA        0        0        0        0        0        0        0
```

3.2. Identificación y tratamiento de valores extremos

El siguiente paso será comprobar que nuestras variables numéricas no contienen valores extremos, y en el caso de que existan decidir que hacer con ellos. Son variables numérciso la edad (Age), el precio del billete (Fare), el número de parientes y esposas (sibsp) y el número de padres e hijos (parch).

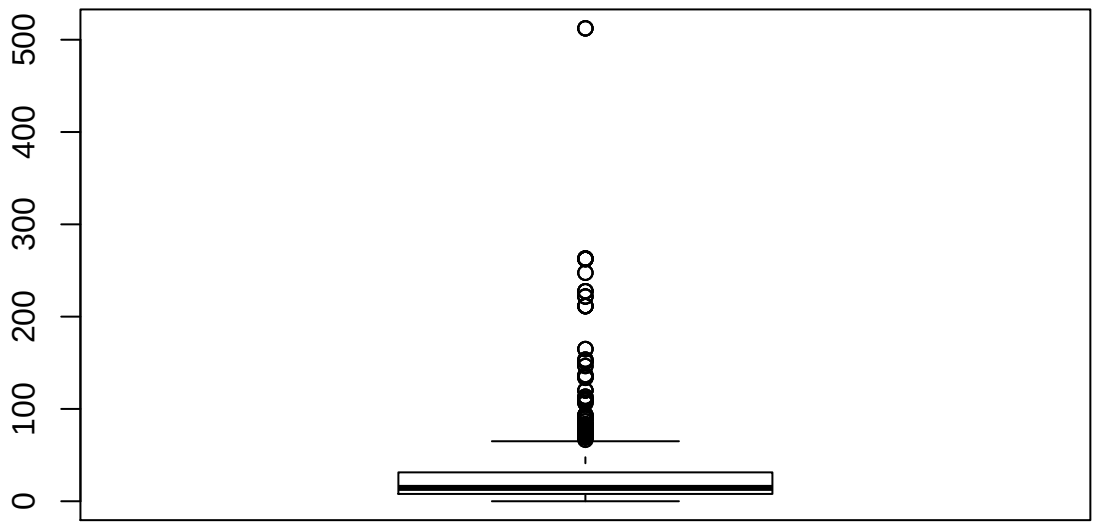
```
#Graficación del boxplot para observar los valores extremos,
#seguidos de la presentación numérica de esos valores
boxplot(full_data$Age)
```




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

```
##  [1] 66.0 65.0 71.0 70.5 65.0 64.0 65.0 71.0 64.0 80.0 70.0 70.0 74.0 67.0 76.0
## [16] 64.0 64.0 64.0
```

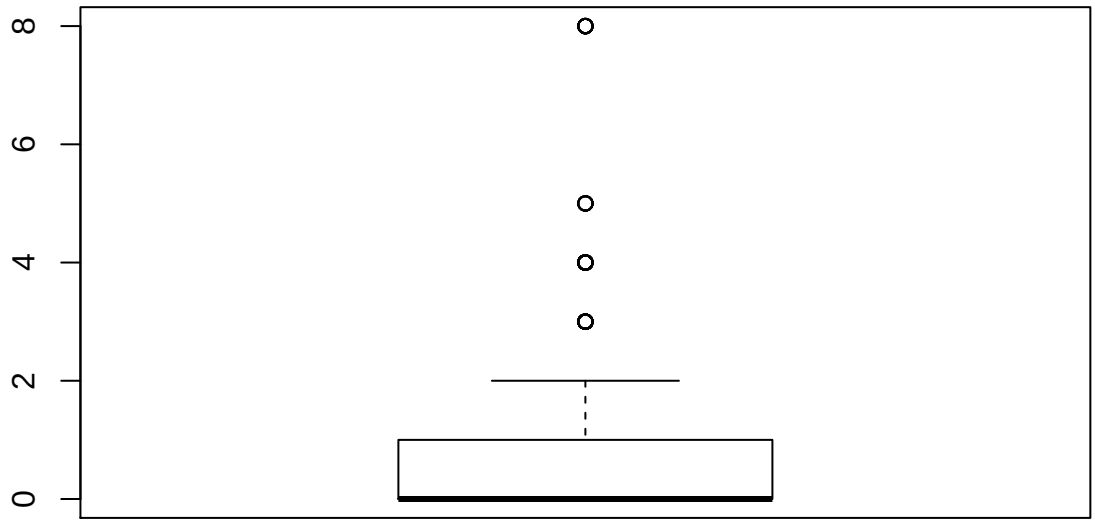
```
#Graficación del boxplot para observar los valores extremos,
#seguidos de la presentación numérica de esos valores
boxplot(full_data$Fare)
```




```
boxplot.stats(full_data$Fare)$out
```

```
##   [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000
##   [9] 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500
##  [17]  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000
##  [25]  79.2000  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500
##  [33]  91.0792 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750
##  [41] 164.8667 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000
##  [49] 263.0000  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000
##  [57] 120.0000 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792
##  [65]  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250
##  [73]  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000
##  [81]  78.2667 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500
##  [89]  73.5000 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250
##  [97] 211.3375 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583
## [105] 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500
## [113]  89.1042 164.8667  69.5500  83.1583  82.2667 262.3750  76.2917 263.0000
## [121] 262.3750 262.3750 263.0000 211.5000 211.5000 221.7792  78.8500 221.7792
## [129]  75.2417 151.5500 262.3750  83.1583 221.7792  83.1583  83.1583 247.5208
## [137]  69.5500 134.5000 227.5250  73.5000 164.8667 211.5000  71.2833  75.2500
## [145] 106.4250 134.5000 136.7792  75.2417 136.7792  82.2667  81.8583 151.5500
## [153]  93.5000 135.6333 146.5208 211.3375  79.2000  69.5500 512.3292  73.5000
## [161]  69.5500  69.5500 134.5000  81.8583 262.3750  93.5000  79.2000 164.8667
## [169] 211.5000  90.0000 108.9000
```

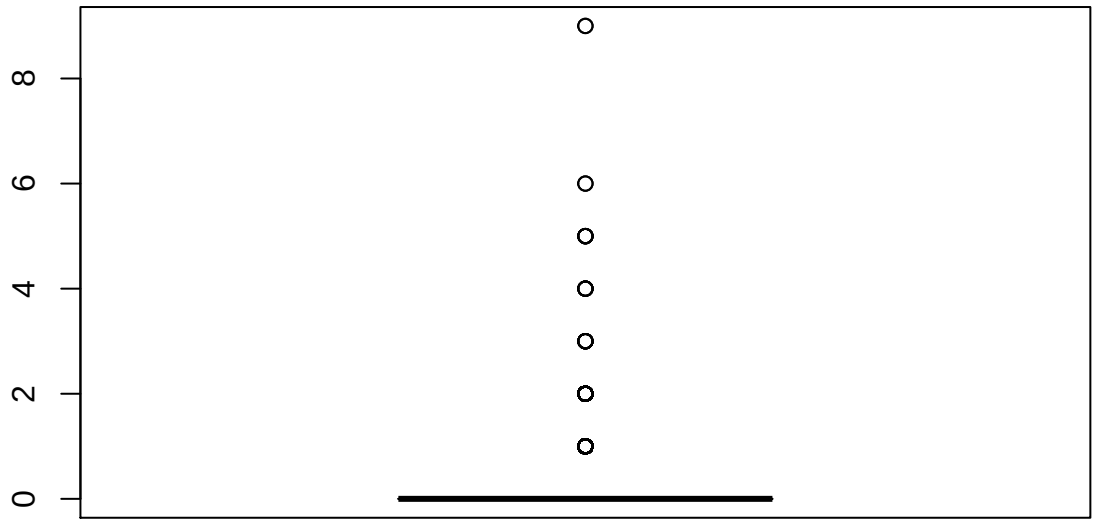
```r
#Graficación del boxplot para observar los valores extremos,
#seguidos de la presentación numérica de esos valores
boxplot(full_data$SibSp)
```




```r
boxplot.stats(full_data$SibSp)$out
```

```
##  [1] 3 4 3 3 4 5 3 4 5 3 3 4 8 4 4 3 8 4 8 3 4 4 4 4 8 3 3 5 3 5 3 4 4 3 3 5 4 3
## [39] 4 8 4 3 4 8 4 8 3 4 5 3 4 8 4 8 4 3 3
```

```r
#Graficación del boxplot para observar los valores extremos,
#seguidos de la presentación numérica de esos valores
boxplot(full_data$Parch)
```




```r
boxplot.stats(full_data$Parch)$out
```

```
##   [1] 1 2 1 5 1 1 5 2 2 1 1 2 2 2 1 2 2 2 3 2 2 1 1 1 1 2 1 1 2 2 1 2 2 2 1 2 1
##  [38] 1 2 1 4 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 1 1 1 1 1 2 1 2
##  [75] 2 1 1 2 1 1 2 1 1 1 1 2 1 1 1 4 1 1 2 2 2 2 2 1 1 1 2 2 1 1 2 2 3 4 1 2 1
## [112] 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 2 2 2 2 2 1 1 2 1 4 1 1 2 1 2 1 1 2 5 2 1 1
## [149] 1 2 1 5 2 1 1 1 2 1 6 1 2 1 2 1 1 1 1 1 1 1 3 2 1 1 1 1 2 1 2 3 1 2 1 2 2
## [186] 1 1 2 1 2 1 2 1 1 1 2 1 1 2 1 2 1 1 1 1 3 2 1 1 1 1 5 2 1 1 1 1 3 1 2 2 1
## [223] 2 1 2 1 2 4 1 1 2 1 1 1 4 6 2 3 1 1 2 2 2 1 1 2 5 2 3 2 1 1 1 2 1 2 2 2 1
## [260] 2 1 1 2 1 2 1 2 1 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2 1 2 9 1 1 1 2 2 2 1 9 1 1
## [297] 2 2 1 1 2 1 1 1 1 1 1
```

Observando los valores extremos que tenemos en nuestros datos, las variables estudiadas y la distribución de los demás valores, se han decidido mantente intactos tal cual están.
En el caso de age, debido a que estos valores extremos son producidos debido a que la distribución de edades indica que la mayor parte de los pasajeros eran jóvenes (mediana alrededor de 30 también).
En el caso del precio del pasaje, los valores extremos se deben a que la mayor parte de billetes han sido de clase baja, como se puede ver en el summary del priemr apartado, centrándonos en los valores de mediana de Pclass.
En el caso de los familiares y la relación padre-hijo, mantendremos también los valores extremos ya que cuadran con la densidad de las familias de la época, siendo 7-8 hijos normal. Para el caso de parentesco, un valor elevado nos puede indicar el viaje de familais enteras, o núcleos familiares grandes.

3.3. Exportación de datasets

```
#Genero los datasets para al regresión
Datos_regresion=na.omit(full_data)
set.seed(123)
train_ind = sample(seq_len(nrow((Datos_regresion))),size = 700)
Datos_regresion_train=Datos_regresion[train_ind,]
Datos_regresion_test=Datos_regresion[-train_ind,]
Datos_predict_kaggle=full_data[is.na(full_data$Survived),]
#Salvo los datasets
write.csv(full_data, "../datasets/titanic_clean_completo.csv", row.names=FALSE)
write.csv(Datos_regresion_train, "../datasets/titanic_clean_regresion_train.csv", row.names=FALSE)
write.csv(Datos_regresion_test, "../datasets/titanic_clean_regresion_test.csv", row.names=FALSE)
write.csv(Datos_predict_kaggle, "../datasets/titanic_predict_kaggle.csv", row.names=FALSE)
```

4. Análisis de los datos

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

A la hora de realizar el análisis, tenemos varias vertientes. La primera sería si existe relación entre el precio pagado y la edad de los viajeros.
La segunda si existe relación entre la clase de viaje y el precio.
La tercera sería ver el impacto del nivel socioeconómico en la supervivencia, es decir, si los ricos se han salvado en su mayor parte frente a los pobres.
También nos gustaría analizar como se distribuyen los diferentes sexos entre los supervivientes y los muertos, y si el puerto de embarque se puede relacionar con haber sobrevivido más o menos.
La cuarta esta relacionada con probar la hipótesis de si la mayor edad va unida a mayor mortalidad o no. Por último queremos intentar predecir si alguien sobrevive o no en función de las variables disponibles Indicar que los análisis por sexo y edad serían apra comprobar parcialmente una frase que es muy mítica en las películas, 'Las mujeres y los niños primero'. Para ello generamos los distintos datasets, que podrían ser utilizados para el análisis.

```
#Datos completos
full_data=read.csv("../datasets/titanic_clean_completo.csv")
#Datos para la parte de regresión
Datos_regresion_train=read.csv("../datasets/titanic_clean_regresion_train.csv")
Datos_regresion_test=read.csv("../datasets/titanic_clean_regresion_test.csv")
Datos_predict_kaggle=read.csv("../datasets/titanic_predict_kaggle.csv")

#Agrupación por Pclass (clase en la que se ha viajado)
full_data.pobre <- full_data[full_data$Pclass == 3,]
full_data.medio <- full_data[full_data$Pclass == 2,]
full_data.rico <- full_data[full_data$Pclass == 1,]

#Agrupación por sexo
full_data.mujer <- full_data[full_data$Sex == 'female',]
full_data.hombre <- full_data[full_data$Sex == 'male',]

#Clases supervivencia por edad
full_data.muertos_edad=full_data[full_data$Survived == 0,]
full_data.muertos_edad=full_data.muertos_edad$Age
full_data.vivos_edad=full_data[full_data$Survived == 1,]
full_data.vivos_edad=full_data.vivos_edad$Age
```

Indicar que estos son algunos de los grupos de interés detectados en los datos, pero que no todos han sido utilizados, o utilizados en forma de datasets distintos durante los análisis,y que en la mayoría de los casos se han extraído directamente del dataset de referencia limpio.

4.2. Comprobación de la normalidad y homogeneidad de la varianza

Para la comprobaciónd e la normalidad de nuestras variables numéricas haremos uso del conocido test de Shapiro-Wilk, en contraposición con otros métodos como Kolmogorov-Smirnov, debido a la robusted y donde comprobaremos que el p-valor es superior a 0.05 para aceptar la hipótesis de normalidad.

```
#Comprobación de la nomrlidad
shapiro.test(full_data$Age)
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  full_data$Age
## W = 0.97671, p-value = 1.066e-13
```

```
shapiro.test(full_data$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  full_data$Fare
## W = 0.52765, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  full_data$SibSp
## W = 0.51108, p-value < 2.2e-16
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  full_data$Parch
## W = 0.49797, p-value < 2.2e-16
```

En ninguno de los casos, el p-valor ha sido superior a 0.05, y por lo tanto debemos suponer que ninguna de nuestras 4 variables numéricas que distribuye normalmente.
Por contra si hiciesemos caso al teorema central del límite, este nos indica que conforme mayor se hace la muestra, podemos considerar que los datos siguen una distribución cuasi normal.

El siguiente paso será comprobar la homocedasticidad.
Al observar el resultado del test de Shapiro-Wilk, debemos aplciar un test no paramétrico de la homocedasticidad, seleccionando el test de Fligner-Killeen para este propósito. En este caso, si tenemos p-valor por encima de 0.05, consideramos igualdad en las varianzas. Este test ya quita los valores NA por defecto, por lo que no nos veremos influídos por aquellos casos en los que no tenemos la variable Survived.

```
#Comprobación de la homocedasticidad
fligner.test(Age ~ Survived, data = (full_data))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Survived
## Fligner-Killeen:med chi-squared = 4.0998, df = 1, p-value = 0.04289
```

```
fligner.test(Fare ~ Survived, data = (full_data))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 96.253, df = 1, p-value < 2.2e-16
```

```
fligner.test(Fare ~ Pclass, data = (full_data))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
```

```
##
## data:  Fare by Pclass
## Fligner-Killeen:med chi-squared = 552.24, df = 2, p-value < 2.2e-16
```

Como podemos ver, en ninguno de los dos casos de interés se cumple la condición de homocedasticidad antendiendo a sus p-valores.

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

4.3.1 Relación entre la edad y el precio del billete

Para este caso aplicaremos la correlación entre ambas variables numéricas. Al estar antes dos variables que no cumplen la condición de normalidad, debemos usar la correlación de Spearman, ya que a diferencia de Pearson, este test no asume normalidad.

```
#Cálculo correlación
test=cor.test(full_data$Age,full_data$Fare,method='spearman', exact=F)
test$estimate
```

```
##       rho
## 0.1806056
```

```
test$p.value
```

```
## [1] 4.632642e-11
```

El resultado nos aporta un alto nivel de confianza gracias al p-valor que obtenemos, pero vemos que el grado de correlación entre las varaibles es muy pequeño.

4.3.2 Relación entre la clase y el precio del billete.¿No varía según clases?

Para este caso aplicaremos el test de Kruskal-Wallis, que vendría a ser la alternativa al ANOVA, que compara una variable cuantitativa con una variable objetivo con más de dos clases.

```
#Cálculo correlación
test=kruskal.test(Pclass ~ Fare, data = full_data)
test$p.value
```

```
## [1] 5.933198e-128
```

Cómo se puede ver en el resultado por el p-valor, hay diferencias significativas entre el precio de las tres clases.

4.3.3 Influencia de las variables cualitativas en la supervivencia

Para buscar estas relaciones, aplicaremos el test de chi cuadrado donde poder ver la relación de las diferentes clases a la hora de viajar, el sexo o dónde se ha embarcado con la supervivencia, ya que estaríamos ante una comparación de variables cualitativas.

```
#Influencia clase sobre superviencia
test=chisq.test(full_data$Survived,full_data$Pclass)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  full_data$Survived and full_data$Pclass
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

```
test$residuals
```

```
##                   full_data$Pclass
## full_data$Survived         1         2         3
##                  0 -4.601993 -1.537771  3.993703
##                  1  5.830678  1.948340 -5.059981
```

En este primer caso observamos como los pasajeros de tercera tienen una relación positiva con haber muerto, es decir, que viajar en tercera clase se relaciona con no haber sobrevivido en mayor medida que las otras dos clases, mientras que haber viajado en primera clase,se relaciona fuertemente con haber sobrevivido.
Estos resultados obtenido los podemos considerar significativos debido a su p-valor menor a 0.05.

```
#Influencia sexo sobre supervivencia
test=chisq.test(full_data$Survived,full_data$Sex)
test
```

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

```
test$residuals
```

```
##                   full_data$Sex
## full_data$Survived    female      male
##                  0 -8.086170  5.965128
##                  1 10.245095 -7.557757
```

Como podemos observar analizando los residuos, ser mujer influye muy positivamente para ser asignado al grupo de supervivientes, mientras siendo hombre ocurre al revés.
Estos resultados obtenido los podemos considerar significativos debido a su p-valor menor a 0.05.

```
#Influencia puerto de embarque sobre supervivencia
test=chisq.test(full_data$Survived,full_data$Embarked)
test
```

```
##
##  Pearson's Chi-squared test
##
## data:  full_data$Survived and full_data$Embarked
## X-squared = 28.005, df = 2, p-value = 8.294e-07
```

```
test$residuals
```

```
##                   full_data$Embarked
## full_data$Survived          C          Q          S
##                  0 -2.90655352 -0.06452452  1.51565538
##                  1  3.68257366  0.08175191 -1.92031991
```

En este caso podemos ver como los embarcados en Southampton están más relacionados con la muerte que los demás puertos (ignorar los valores vacíos, cosa que hacemos solo para este caso). Esto podría deberse a que este fue el puerto de salida, y por lo tanto, que la tripulación, últimos en abandonar el barco, se embarcase en ese puerto resultando en los valores que tenemos.
Estos resultados obtenido los podemos considerar significativos debido a su p-valor menor a 0.05.
Indicar finalmente, que por como funciona el test de chi cuadrado, podemos seguir trabajando con el full_data, pese a tener valores NA en la supervivencia, ya que estas entradas son automáticamente eliminadas. Esta decisión se ha tomado teniendo en consideración como funciona el algoritmo y para mantener lo máximo posible full_data para los demás análisis.

4.3.4 ¿La supervivencia es menor en función de la edad del pasajero?

La tercera prueba estadística consiste en aplicar el contraste de hipótesis sobre dos muestras para determinar si una mayor edad está ligada a no haber sobrevivido. Estaríamos comparando una variable cuantitiva como es al edad con una cualitativa de 2 niveles que es la supervivencia. Ya que tenemos una n>30 podemos utilizar un método paramétrico para este contraste, siguiendo el teorema central del límite.

```
#Comprobación de hipótesis
test=t.test(full_data.muertos_edad,full_data.vivos_edad,paired=FALSE, alternative="greater")
test
```

```
##
##  Welch Two Sample t-test
##
## data:  full_data.muertos_edad and full_data.vivos_edad
## t = 2.5446, df = 686.65, p-value = 0.005579
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.874927      Inf
## sample estimates:
## mean of x mean of y
##  30.20765  27.72711
```

Como tenemos un p_valor mucho menor a 0.05 podemos rechazar la hipótesis nula, concluyendo que a mayor edad, menor será la supervivencia.

4.3.5 Modelo de Regresión Logística

En nuestro caso, nos interesa predecir la variable objetivo Supervivencia, por lo que debemos alpicar un modelo de regresión que nos clasifique en las dos clases. Por ello vamos a hacer uso de los generalized linear models en su configuración binaria para crear 5 modelos de regresión logística. Acto seguido usaremos el mejor modelo para predecir la supervivencia del conjunto de test.

```
#Regresores cuantitativos
Age=Datos_regresion_train$Age
SibSp=Datos_regresion_train$SibSp
Parch=Datos_regresion_train$Parch
Fare=Datos_regresion_train$Fare

#Regresores cualitativos
Pclass=Datos_regresion_train$Pclass
Sex=Datos_regresion_train$Sex
Embarked=Datos_regresion_train$Embarked
#Variable objetivo
Survived=Datos_regresion_train$Survived
#Modelos
modelo1 <- glm(Survived ~ Age + SibSp + Parch + Fare,
               data = Datos_regresion_train, family = "binomial")
AIC(modelo1)
```

```
## [1] 869.7876
```

```
modelo2 <- glm(Survived ~ Pclass + Sex + Embarked,
               data = Datos_regresion_train, family = "binomial")
AIC(modelo2)
```

```
## [1] 658.0027
```

```
modelo3 <- glm(Survived ~ Age + Pclass + Sex ,
               data = Datos_regresion_train, family = "binomial")
AIC(modelo3)
```

```
## [1] 641.9972
```

```
modelo4 <- glm(Survived ~ Age + Pclass + Sex + Embarked,
               data = Datos_regresion_train, family = "binomial")
AIC(modelo4)
```

```
## [1] 637.038
```

```
modelo5 <- glm(Survived ~ Age + Fare + Sex ,
               data = Datos_regresion_train, family = "binomial")
AIC(modelo5)
```

```
## [1] 705.6571
```

```
modelo6 <- glm(Survived ~ Age + Pclass + Sex + SibSp + Embarked,
               data = Datos_regresion_train, family = "binomial")
AIC(modelo6)
```

```
## [1] 623.4662
```

De los modelos probados nos quedamos con el que presenta menor valor de AIC que valora como se adapta el modelo a los datos, pero penalizando el número de variables que se introducen.

```
#Predict sobre el pequeño conjunto marcado que nos hemos guardado

prediction=predict(modelo6,
                   newdata=subset(Datos_regresion_test,select=c(2,3,4,5,8)),
                   type='response')
prediction <- ifelse(prediction > 0.5,1,0)
prediction
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   0   1   0   1   1   0   0   0   0   1   0   1   0   1   0   0   0   0   0   0
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   0   1   0   1   0   0   0   1   1   0   0   0   0   0   0   0   1   0   0   0
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   0   0   0   0   0   1   0   0   1   1   0   0   1   0   0   0   0   0   0   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   1   0   0   1   0   0   0   0   1   0   0   1   0   0   0   0   1   0   0   0
##  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
##   1   1   0   1   0   0   0   1   1   1   1   1   0   0   0   0   0   0   0   0
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
##   0   1   0   0   0   1   1   0   0   0   1   1   1   0   0   1   1   0   0   1
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
##   1   0   1   1   0   1   0   0   0   0   1   0   0   0   0   1   0   1   1   1
## 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
##   1   0   0   0   0   0   1   1   0   0   0   0   0   0   1   0   0   1   0   0
## 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   0   1   0   1   0   0   1   0   0   1   0   1   0   0   0   0   0   0   0   1
## 181 182 183 184 185 186 187 188 189 190 191
##   0   1   0   0   1   1   0   0   1   0   0
```

```
misClasificError <- mean(prediction != Datos_regresion_test$Survived)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.785340314136126"
```

Una vez hemos probado nuestro modelo en datos conocidos y hemos obtenido un 78.5% de acierto, podemos aplicar este modelo a nuevos datos, en este caso a los Datos_predict, que corresponden con el dataset de prueba de kaggle (test.csv), pero tras haber imputado valores nulos.

```
#Predict sobre el pequeño conjunto marcado que nos hemos guardado
prediction_kaggle=predict(modelo6,
                          newdata=subset(Datos_predict_kaggle,select=c(2,3,4,5,8)),
                          type='response')
prediction_kaggle <- ifelse(prediction_kaggle > 0.5,1,0)
prediction_kaggle
```

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
##   0   0   0   0   1   0   1   0   1   0   0   0   1   0   1   1   0   0   0   0
##  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
##   0   0   1   1   1   0   1   0   0   0   0   0   0   0   1   0   1   1   0   0
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60
##   0   0   0   1   1   0   0   0   1   0   0   0   1   1   0   0   0   0   0   1
##  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80
##   0   0   0   1   1   1   1   0   1   1   1   0   1   1   1   1   0   1   0   1
##  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
##   0   0   0   0   0   0   1   1   1   0   1   0   1   0   1   0   1   0   1   0
## 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
##   1   0   0   0   1   0   0   0   0   0   0   1   1   1   1   0   0   1   1   1
## 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
##   1   0   1   0   0   1   0   1   0   0   0   0   0   0   0   0   0   0   1   0
## 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
##   0   1   0   0   0   0   0   0   0   0   1   0   0   0   0   0   1   1   0   1
## 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
##   1   0   1   0   0   0   0   0   1   1   0   0   0   0   0   1   1   0   1   1
## 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
##   0   0   1   0   1   0   1   0   0   0   0   0   0   0   0   0   1   1   0   1
## 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220
##   1   0   0   1   0   0   1   0   1   0   0   0   0   0   0   0   1   0   1   0
## 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
##   1   0   1   0   1   1   0   1   0   0   0   1   0   0   0   0   0   0   1   1
## 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260
##   1   1   0   0   0   0   1   0   1   1   1   0   1   0   0   0   0   0   1   0
## 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280
##   0   0   1   1   0   0   0   0   1   0   0   0   1   1   0   1   0   0   0   0
## 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300
##   1   0   1   1   1   0   0   0   0   0   0   1   0   0   0   0   1   0   1   0
## 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320
##   0   0   0   0   1   1   0   0   0   0   0   0   0   1   1   1   0   0   0   0
## 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340
##   0   0   0   0   1   0   1   0   0   0   1   1   0   1   0   1   0   0   0   0
## 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360
##   0   0   0   1   0   1   0   1   0   1   1   0   0   0   1   0   1   0   0   0
## 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380
##   0   1   1   0   1   0   0   1   1   0   0   1   0   0   1   1   0   0   0   0
## 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400
##   0   0   1   1   0   1   0   0   0   0   1   1   0   0   0   1   0   1   0   0
## 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418
##   1   0   1   1   0   0   0   0   1   1   1   1   1   0   1   0   0   0
```

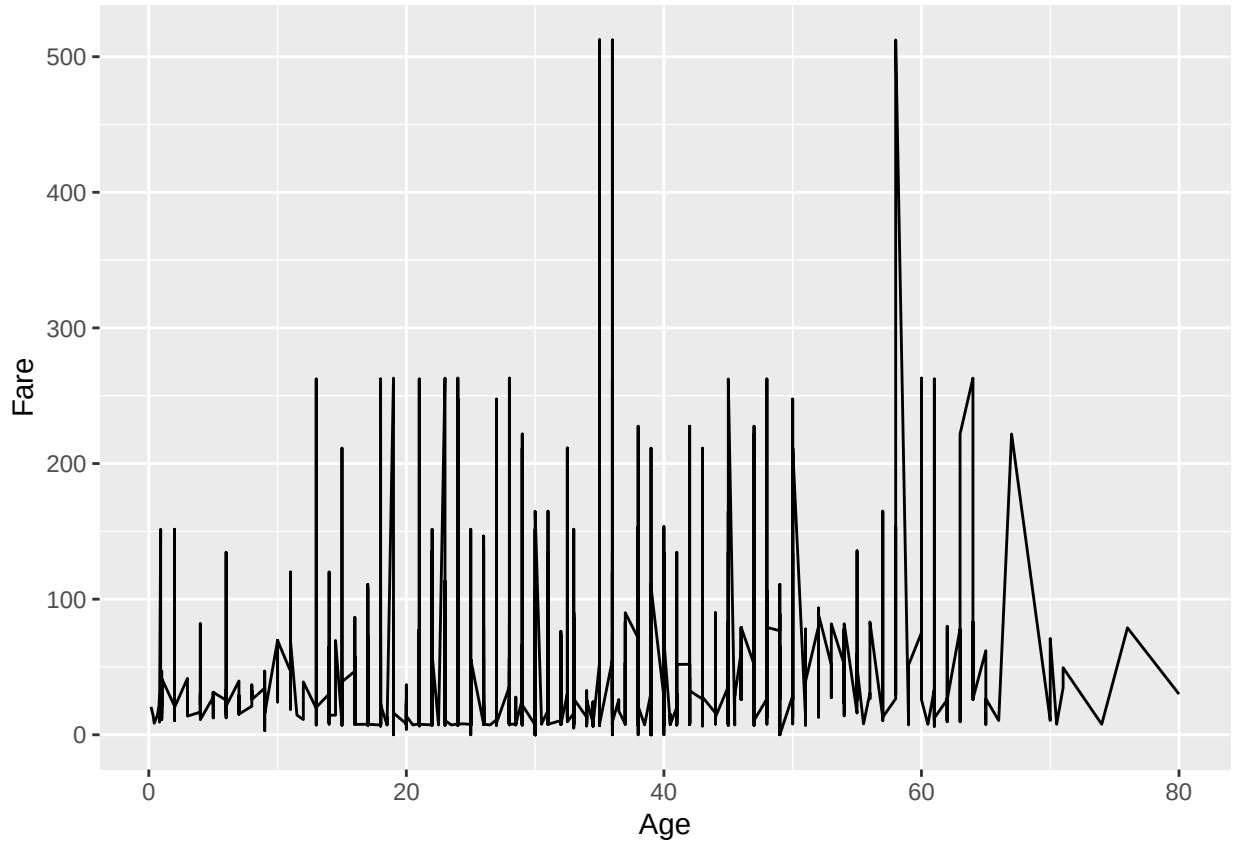
5. Representación de los resultados a partir de tablas y gráficas

Realizaremos la representación de los resultados de los cinco objetivos de análisis resultantes de comparar grupos de datos.
Las representaciones aquí presentes vendrán a complementar las ya existentes en el apartado de análisis, que en su mayor parte han sido tabla y resultados de la ejecución de los distintos algoritmos.

5.1 Relación entre la edad y el precio del billete

En este apartado hemos visto como estadísticamente no había relación entre el precio del billete y la edad de los pasajeros. Vamos a realizar una representación de estas dos variables mediante una gráfica lineal con todas las filas para observar esta ausencia de relación.

```
#Código de representación lineal
library(ggplot2)
ggplot(full_data,aes(x=Age, y=Fare)) + geom_line()
```

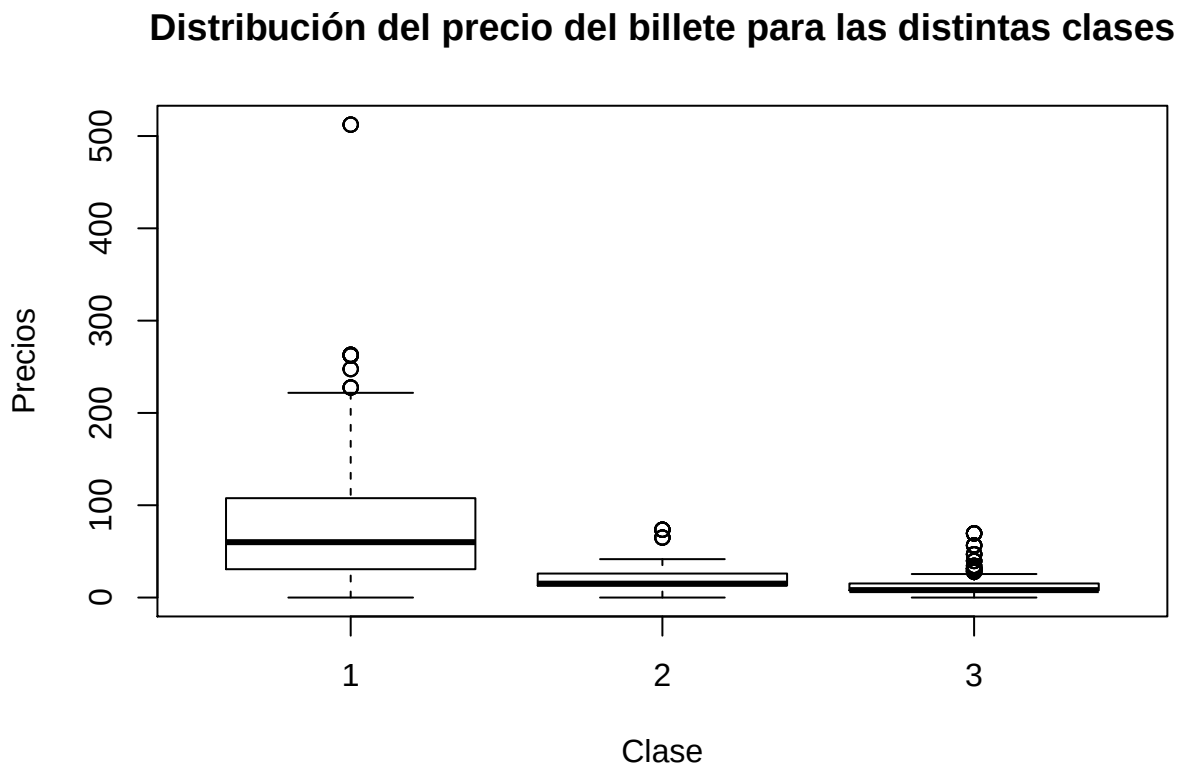


En la gráfica podemos ver dos picos, pero también como los valores de Fare son independientes de la edad teniendo valores muy similares en muchos puntos del espectro de edad.

5.2 Relación entre la clase y el precio del billete.¿No varía según clases?

Para este caso para a realizar una visualización con boxplot que nos permita ver como se distribuye el precio del billete en las distintas clases, e intentar apoyar los resultados de la parte de análisis, que indicaban diferencias en el precio según clase.

```
#Código de representación en caja de la distribución
boxplot(Fare~Pclass,data=full_data,
        main="Distribución del precio del billete para las distintas clases",
   xlab="Clase", ylab="Precios")
```

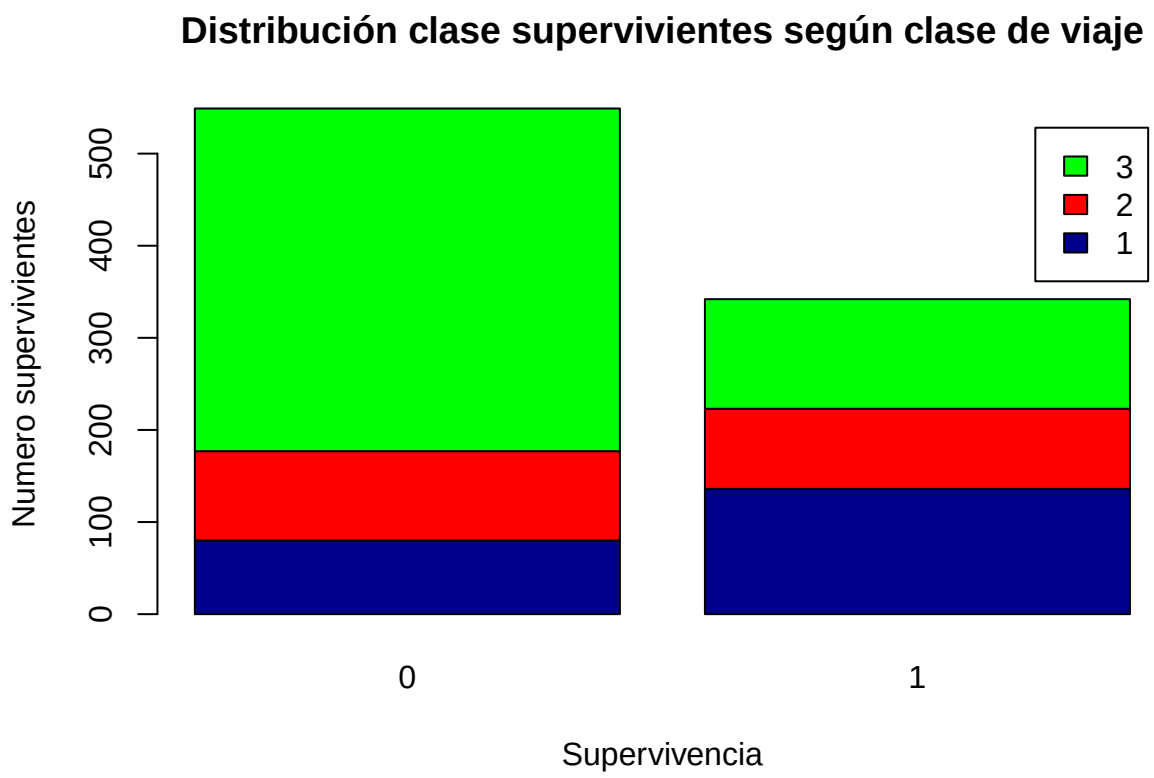


Donde podemos observar como los precios para la clase más alta (1) son bastante superiores a las otras dos, habiendo pequeña diferencia entre las 2 y 3.

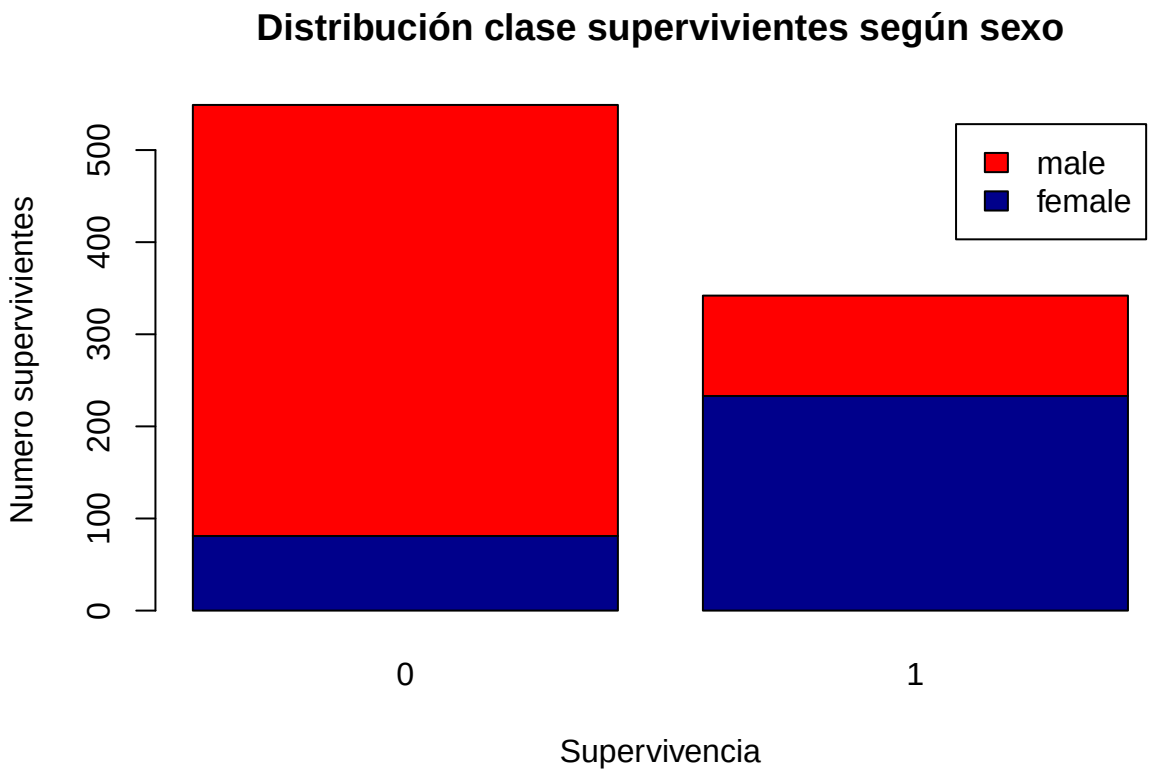
5.3 Influencia de las variables cualitativas en la supervivencia

En las siguientes tres gráficas de barras venimos a comprobar los resultados obtenidos por los test de la chi cuadrado que nos mostraban las relaciones entre muestras. Ahora somos capaces de ver cada variable distribuida dentro del número de supervivientes y niveles de la clase superviviente, con el fin de tener un apoyo gráfico a la tablas y resultados obtenidos en el apartado 4.3.3.

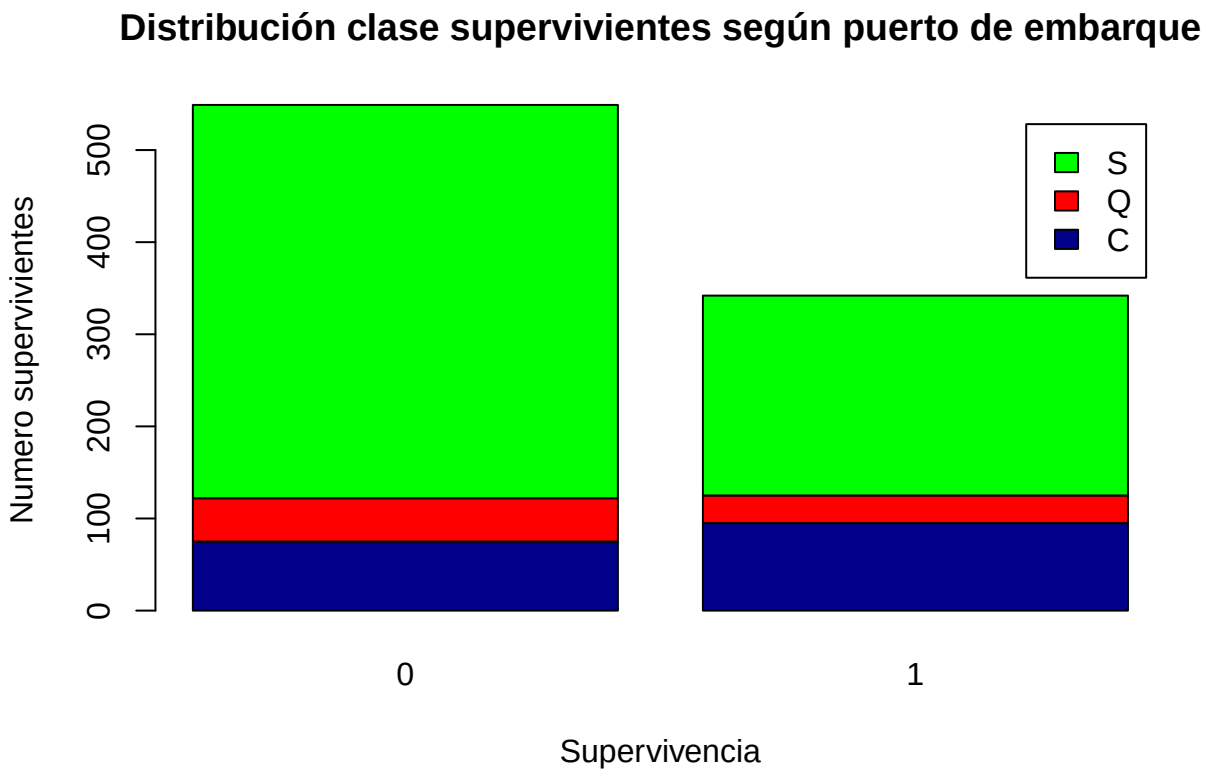
```
#Código de representación barras
counts <- table(full_data$Pclass,full_data$Survived)
barplot(counts, main="Distribución clase supervivientes según clase de viaje",
  xlab="Supervivencia", ylab='Numero supervivientes', col=c("darkblue","red",'green'),
  legend = rownames(counts))
```



```r
#Código de representación barras
counts <- table(full_data$Sex,full_data$Survived)
barplot(counts, main="Distribución clase supervivientes según sexo",
  xlab="Supervivencia", ylab='Numero supervivientes', col=c("darkblue","red"),
  legend = rownames(counts))
```

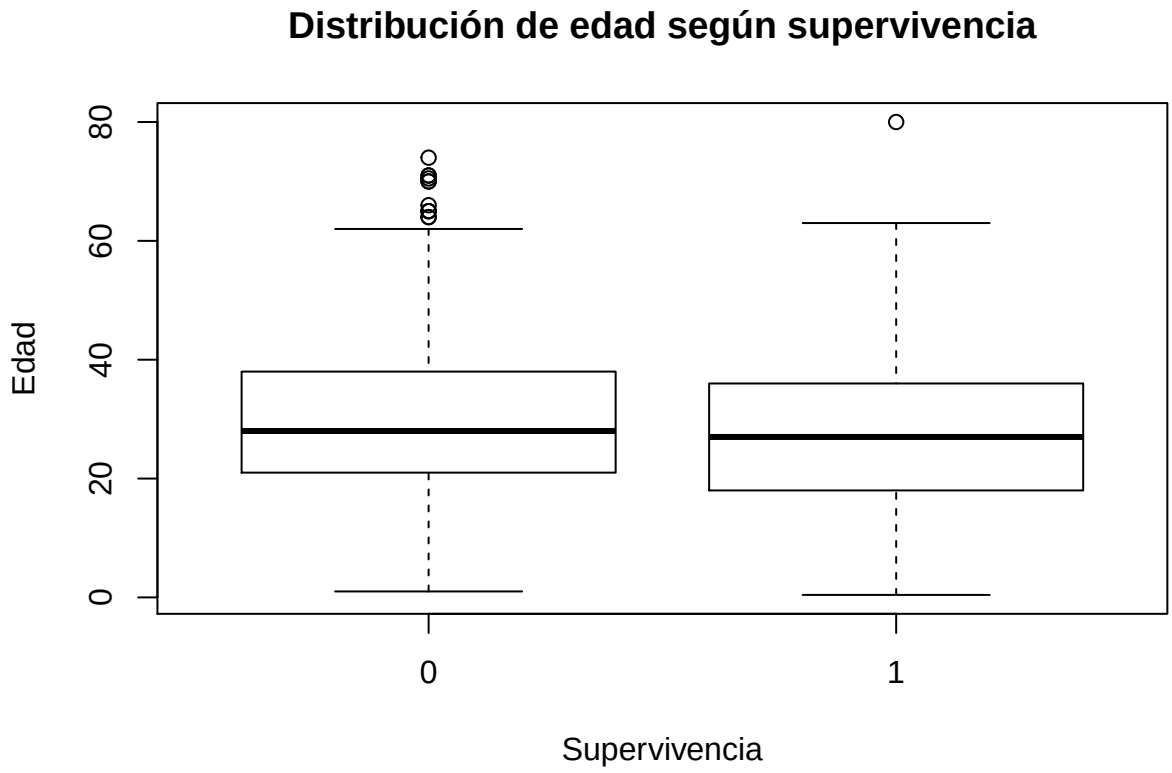


```r
#Código de representación barras
counts <- table(full_data$Embarked,full_data$Survived)
barplot(counts, main="Distribución clase supervivientes según puerto de embarque",
  xlab="Supervivencia", ylab='Numero supervivientes', col=c("darkblue","red",'green'),
  legend = rownames(counts))
```



5.4 ¿La supervivencia es menor en función de la edad del pasajero?

```
#Código de representación en caja de la distribución
boxplot(Age~Survived,data=full_data,
        main="Distribución de edad según supervivencia",
   xlab="Supervivencia", ylab="Edad")
```



Esta representación viene a indicar lo mismo que en la parte del análisis, pero ahora también podemos observar que las diferencias existen pero son pequeñas(en análisis eran 30 años vs 27 aproximadamente, similar a lo que se nos indica aquí).

5.5 Modelo de Regresión Logística

Todas las representaciones necesarias se han hecho mediante tablas en el apartado de análisis, tanto para la prediction del conjunto de test, como la predicción del conjunto de kaggle, como para los valores de validación de los diferentes modelos.

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

A lo largo de este trabajo hemos realizado la limpieza de los datos del titanic disponibles en la página web kaggle. Hemos seleccionado aquellas variables que podrían ser interesantes para realizar un análisis y hemos realizado el análisis.
Del análisis podemos extraer diferentes cosas, como que ninguna de nuestra variables cuantitativas seguía una distribución normal ni cumplía con las condiciones de homocedasticidad.
En lo referente a las relaciones entre variables, hemos intentado aplicar diversas técnicas siempre teniendo en cuenta los datos con los que estábamos trabajando:

- Cómo primer objeto de análisis se quiso estudiar si la edad de los viajeros influía en el precio que estos paganba por subir a bordo. Del análisis estadístico extrajimos que la relación entra ambas variables era muy pequeña y mediante el análisis visual pudimos ver como el precio pagado por el pasaje no sufría variaciones de distribución significativas conforme avanzamos en la edad.

- Cómo segundo objetivo teníamos el averiguar si el precio del billete era insensible al cambio de clase, algo que mediante el test realizado quedó falsado con un p_valor muy por debajo del límite 0.05. Esto nos indicaba que habías variaciones en el precio según la clase en la que se viajase. Para complementar este análisis y ver como eran esas variaciones, realizamos la graficación boxplot que nos mostró como la primera clase era la más cara con diferencia, siendo la más barata la tercera clase. Al carecer de fechas de compra de los billetes, no podemos ahondar más en nuestro análisis sobre las causas de las fluctuaciones de precio.

- Nuestro tercer objetivo era ver como se relacionaban el sexo, la clase de viaje y el puerto de embarque con la supervivencia. Cómo conclusión, los hombres murieron muchos más que las mujeres, la primera clase daba más garantías de salvarse y haber embarcado en el primer puerto, Southampton, también estaba ligado a mayor mortalidad. Mediante los métodos gráficos del apartado anterior se pueden ver con más claridad estas relaciones y confirmar visualmente los resultados del análisis de chi cuadrado. Los resultados de sexo también pueden ir ligados al personal de tripulación del barco y sería un interesante análisis para futuro.

- En el cuarto apartado de relaciones entre variables, quisimos comprobar la hipótesis en la que se asociaba mayor edad a más mortalidad. Esta hipótesis a quedado comprobada aunque mediante los métodos gráficos hemos visto como esta diferencia entre grupos no es tan elevada como nos pensábamos a priori. Esto se puede deber a la mayor complicación de sobrevivir al naufragio a mayor edad, o a la afirmación de que 'mujeres y niños primero', como también para para el tercer objetivo y el sexo.

- En el quinto apartado, empleamos nuestro tiempo en intentar realizar un modelo predictivo clasificador basado en regresión logística haciendo uso de alguna de las variables disponibles. El mejor modelo de los que probamos fue el que involucraba edad, sexo, clase de viaje, relaciones de parentesco y puerto de embarque. Aplicado a un conjunto de test, hemos obtenido un valor de 78.5% de precisión, el cual no está mal comparado con los valores que aparecen en kaggle, pero que podría ser mejorable hasta alguno casos con aproximadamente 84%, según la competición de kaggle.
  Indicar que también hemos hecho la predición en base a los datos de test de Kaggle, tras haber realizado las impotacones pertinentes en los valores vacios, y que variar el método de imputación puede ser clave para obtener mejores resultados, aplicando por ejemplo la media, pero en este trabajo se ha querido probar la imputación KNN.
  Finalmente cabe decir, que se nos ha quedado en el tintero un análisis muy interesante, los grupos de edad y supervivencia. Para un futuro, y como complemento al contraste de hipótesis que hemos hecho, podríamos discretizar la variable edad. Para ello podríamos tomar menos de 14 años como niños, de 14 a 50 como edad media y mayores de 50 como viejos, ya que la esperanza de vida para inicios del siglo XX se encontraba entre 50 y 65 (https://es.wikipedia.org/wiki/Esperanza_de_vida).

7. Código

Tanto el código, como este mismo archivo, como los conjuntos de datos utilizados están disponibles en https://github.com/Javipercor/Analisis_Titanic-R/