PRÁCTICA 2

```
library("stats")
library("class")
library("VIM")
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

Práctica 2

1. Descripción del dataset.

Hemos decidido analizar el dataset de los pasajeros del Titanic.

El conjunto de datos está formado por 891 filas, además del encabezado, con las siguientes variables:

- survived: Supervivencia, si el registro es un cero nos indica que no sobrevivió y con un 1 sí sobrevivio.
- pclass: clase de ticket. Opciones: 1=Primera, 2= Segunda, 3= Tercera.
- name: Nombre del pasajero.
- sex: Sexo.
- age: Edad.Si la edad es inferior a 1, aparece fraccional. Además, si esta es estimada, aparecerá de la forma xx.5.
- stbsp: Hermanos o conyuges a bordo del Titanic.
- parch: Padre o hijos a bordo del Titanic.
- ticket: Número del ticket.
- fare: Tarifa del pasajero.
- cabin: Número de cabina.
- embarked: Puerto de embarque.

Este conjunto de datos puede responder a las causas de las muertes en el naufragio del Titanic, pudiendo establecer modelos sobre las cuasas relativas a la mortandad entre los diferentes tipos de pasajeros. También puede facilitar un modelo de interés sobre qué variables han influido en las muertes.

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

En primer lugar procedemos a realizar la lectura del fichero `train`, en formato csv. Esto nos devuelve un `data.frame`.

```
datos <- read.csv("all/train.csv", header = TRUE)
```

Hacemos uso de la función `class` para saber qué tipo de datos se asigna por defecto a cada campo del dataset.

```
sapply(datos, 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"
```

Elegimos transformar en tipo factor los campos `Survived` y `Pclass`.

Podemos prescindir, inicialmente, de los campos `PassengerId`, `Name`. Transformamos en factor los campos `Survived`, `Pclass`.

```
datos <- datos[, -c(1, 4)]
datos$Survived <- as.factor(datos$Survived)
datos$Pclass <- as.factor(datos$Pclass)
```

En cuanto a la selección de los datos de interés, en un primer análisis contemplaremos como variables de entrada los campos `Pclass`, `Sex`, `Age` y `Embarked`.

3. Limpieza de los datos.

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

Con la función `is.na()` podemos ver los datos vacíos que contiene nuestro dataset.

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

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

```
summary(datos)
```

```
##  Survived Pclass      Sex           Age            SibSp
##  0:549    1:216   female:314   Min.   : 0.42   Min.   :0.000
##  1:342    2:184   male  :577   1st Qu.:20.12   1st Qu.:0.000
##           3:491                Median :28.00   Median :0.000
##                                Mean   :29.70   Mean   :0.523
##                                3rd Qu.:38.00   3rd Qu.:1.000
##                                Max.   :80.00   Max.   :8.000
##                                NA's   :177
##      Parch             Ticket         Fare               Cabin
##  Min.   :0.0000   1601    :  7   Min.   :  0.00             :687
##  1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91   B96 B98    :  4
##  Median :0.0000   CA. 2343:  7   Median : 14.45   C23 C25 C27:  4
##  Mean   :0.3816   3101295 :  6   Mean   : 32.20   G6         :  4
##  3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00   C22 C26    :  3
```

```
##  Max.   :6.0000   CA 2144 :  6   Max.   :512.33   D          :  3
##                   (Other) :852                    (Other)    :186
##  Embarked
##   :  2
##  C:168
##  Q: 77
##  S:644
##
##
##
```

Teniendo en cuenta estas salidas, podemos decir que:

- `Pclass` está completo.
- En `Age` tenemos NAs. Concretamente 177.
- Tanto `SibSP` como `Parch` tienen valores iguales a cero. Consideramos esto totalmente normal y no requiere análisis, debido a que habrá gran parte de integrantes del pasaje que viajen solos o viajen con alguien pero que no sea de su familia.
- En la variable `Fare` tenemos valores 0. ¿Esto puede significar que hubiera viajeros que viajasen gratis o que no se tuvieran datos de lo que pagaron? Se supone que, siendo un viaje inaugural, hubo invitados. Por lo tanto, estos valores 0 no se tratarán de modo especial.
- Tenemos 687 registros que carecen del dato reflejado en `Cabin`.
- En la variable `Embarked` aparecen dos pasarejos sin clasificar en ninguno de los tres puertos. Estos dos registros los tendremos que tratar.

En este punto, en primer lugar creemos necesario imputar los 177 valores perdidos de la edad. Lo que haremos con la función `kNN` del paquete VIM.

```
datos$Age <- kNN(datos)$Age
summary(datos$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   21.00   28.00   29.70   38.00   80.00
```

Se puede comprobar que sólo hay un pequeño cambio en el primer cuartil con respecto a los datos originales, que pudimos ver en la ejecución del anterior `summary`.

En cuanto a los dos registros en los que falta el campo `Embarked`, podemos observar que en ambos casos, los pasajeros han sido "colocados" en un camarote que empieza por "B". Vamos a comprobar si de aquí podemos sacar algo:

```
datos[datos$Embarked == "",]
```

```
##     Survived Pclass    Sex Age SibSp Parch Ticket Fare Cabin Embarked
## 62         1      1 female  38     0     0 113572   80   B28
## 830        1      1 female  62     0     0 113572   80   B28
```

Una de las cosas que podemos observar es dónde ha embarcado cada pasajero alojado en un camarote de primer caracter "B":

```
table(datos[substring(datos$Cabin, 1, 1) == 'B',]$Embarked)
```

```
##
##     C  Q  S
##  2 22  0 23
```

Vemos que hay prácticamente la misma probabilidad de que hayan embarcado en C o en S y muy poca en Q. Así que, sabiendo esto, y que la probabilidad total de haber embarcado en uno u otro puerto es bastante mayor de haber embarcado en S:

```
table(datos$Embarked)
```

```
##
##       C   Q   S
##   2 168  77 644
```

Asumimos que los dos pasajeros sin puerto, del camarote B28, habrán embarcado en S.

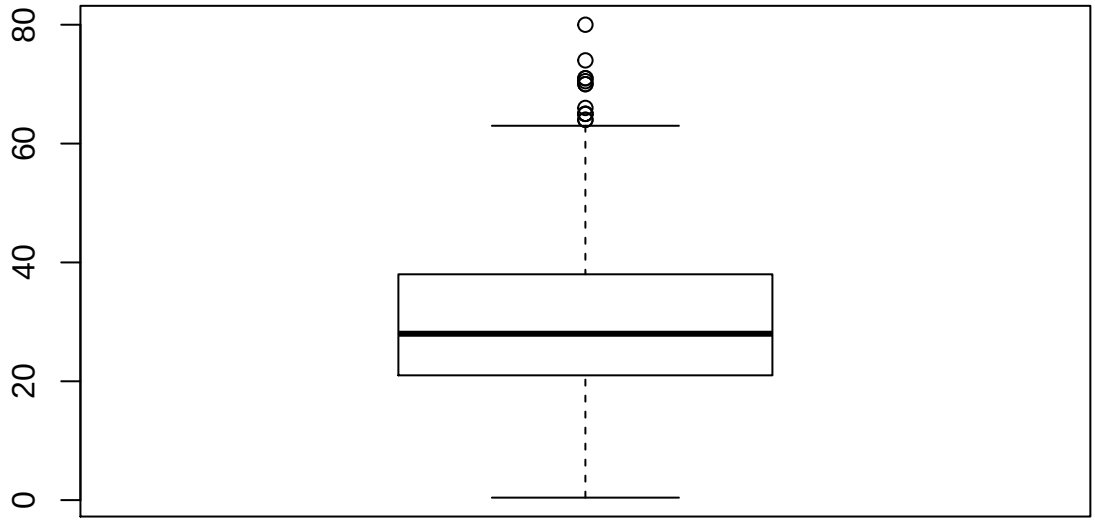
```
datos$Embarked[62] <- as.factor("S")
datos$Embarked[830] <- as.factor("S")
summary(datos$Embarked)
```

```
##       C   Q   S
##   0 168  77 646
```

3.2. Identificación y tratamiento de valores extremos.

En estadística, tales como muestras estratificadas, un valor atípico (outlier) es una observación que es numéricamente distante del resto de los datos. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas. Los valores atípicos pueden ser indicativos de datos que pertenecen a una población diferente del resto de las muestras establecidas. En nuestro caso, el único caso con valore extremos, como podemos ver en la siguiente imagen, es la edad. Ninguno de estos valores afecta significativamente a nuestro estudio.

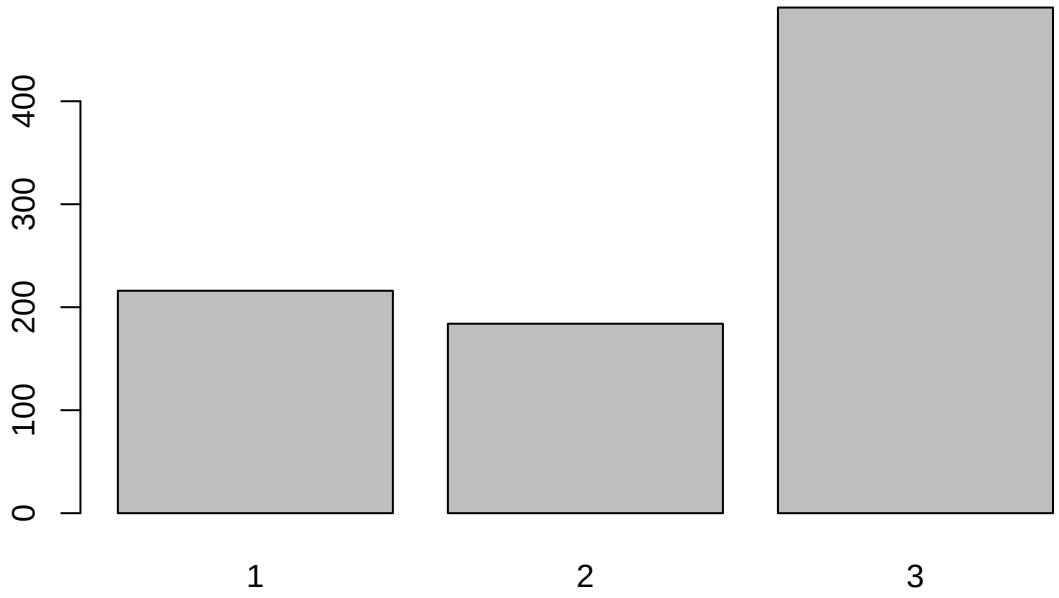
```
boxplot(datos$Age)
```



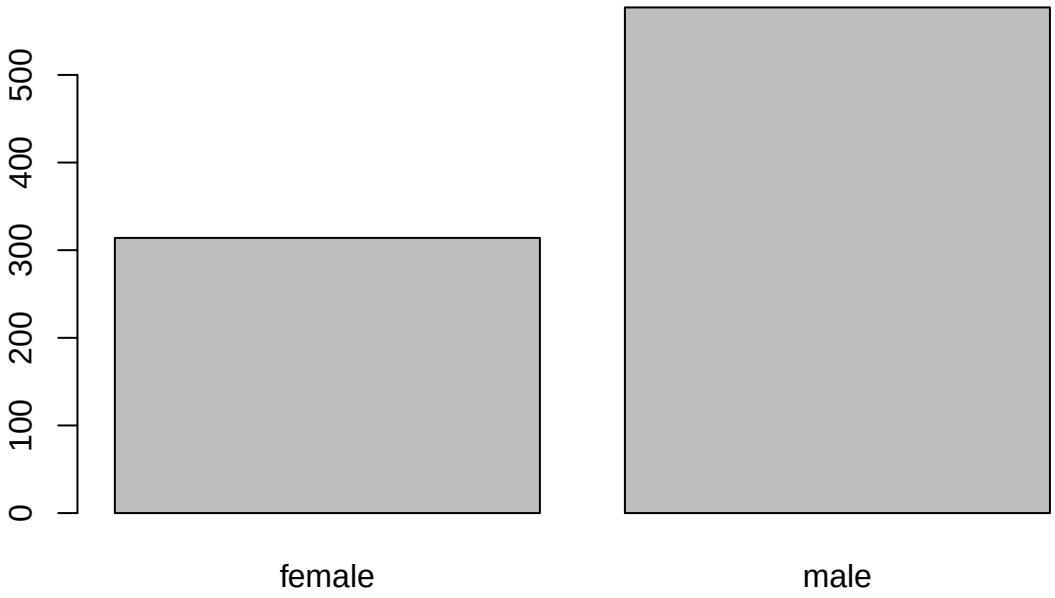
4. Análisis de los datos.

Tras importar y revisar detalladamente todas las variables incluidas en nuestro dataset, hemos decidido hacer el estudio con los siguientes campos: survived, pclass, sex, age, embarked

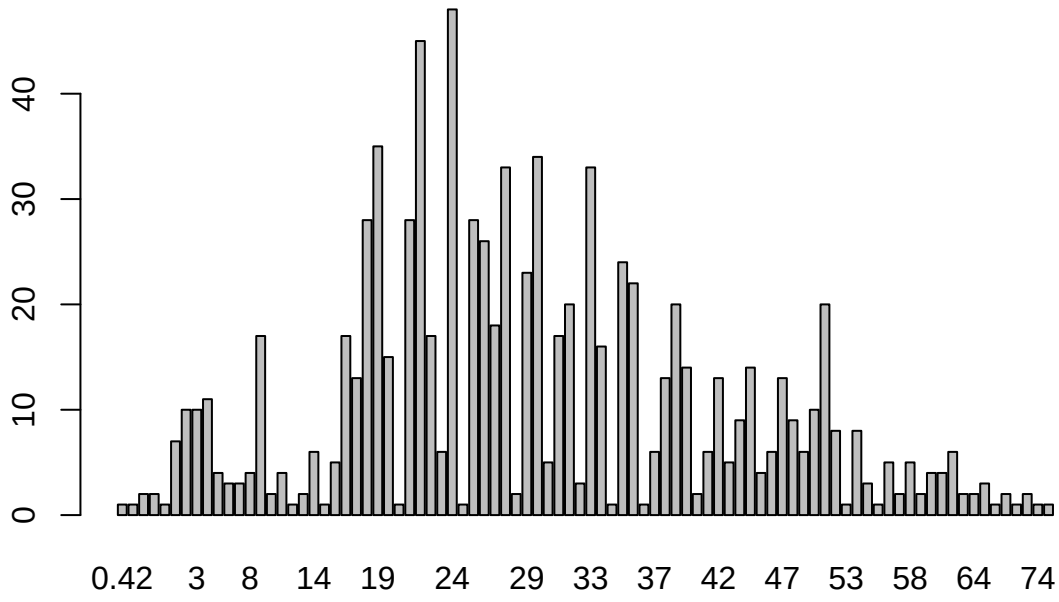
```
barplot(table(datos$Pclass))
```



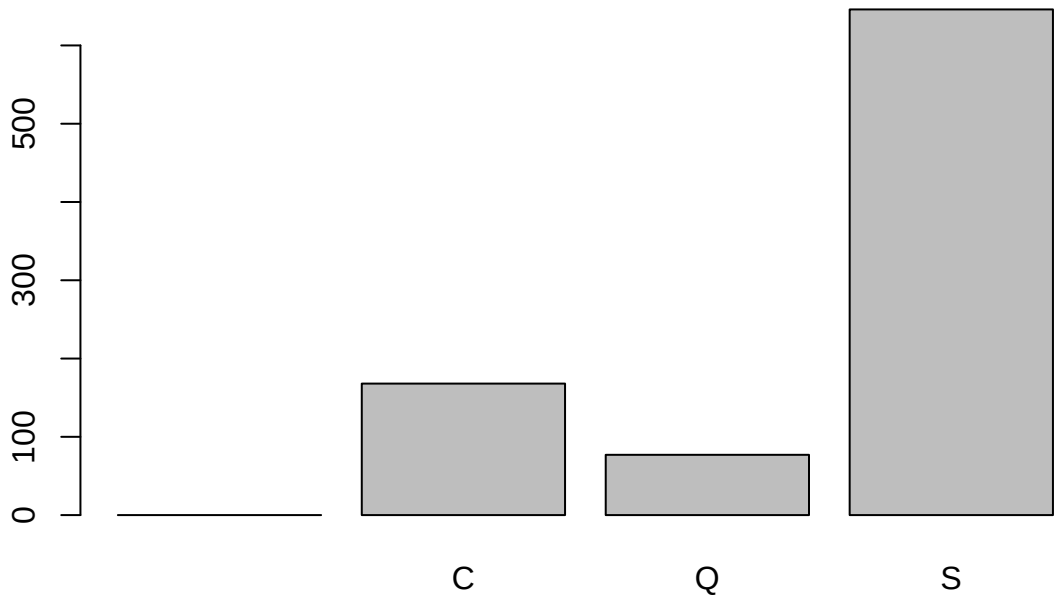

```
barplot(table(datos$Sex))
```




```
barplot(table(datos$Age))
```




```
barplot(table(datos$Embarked))
```



Cuando queremos evaluar el grado de asociación o independencia entre una variable cuantitativa y una variable categórica (y recuérdese que ésta clasifica o diferencia a los individuos en grupos, tantos como categorías tiene dicha variable), el procedimiento estadístico inferencial recurre a comparar las medias de la distribuciones de la variable cuantitativa en los diferentes grupos establecidos por la variable categórica. Si ésta tiene solo dos categorías (es dicotómica), la comparación de medias entre dos grupos independientes se lleva a cabo por el test t de Student; si tiene tres o más categorías, la comparación de medias entre tres o más grupos independientes se realiza a través de un modelo matemático más general, el Análisis de la Varianza (ANOVA).

La distribución normal es una distribución con forma de campana donde las desviaciones estándar sucesivas con respecto a la media establecen valores de referencia para estimar el porcentaje de observaciones de los datos. Estos valores de referencia son la base de muchas pruebas de hipótesis, como las pruebas Z y t.

Como hemos indicado en 4.1, se aplicará en primer lugar un ANOVA a las variables indicadas.

```
a1 <- aov(as.integer(datos$Survived) ~ as.integer(datos$Pclass))
anova(a1)
```

```
## Analysis of Variance Table
##
## Response: as.integer(datos$Survived)
##                            Df  Sum Sq Mean Sq F value    Pr(>F)
## as.integer(datos$Pclass)    1  24.143 24.1429  115.03 < 2.2e-16 ***
## Residuals                 889 186.584  0.2099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
a2 <- aov(as.integer(datos$Survived) ~ as.integer(datos$Sex))
anova(a2)
```

```
## Analysis of Variance Table
##
## Response: as.integer(datos$Survived)
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## as.integer(datos$Sex)    1  62.213  62.213  372.41 < 2.2e-16 ***
## Residuals              889 148.514   0.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
a3 <- aov(as.integer(datos$Survived) ~ as.numeric(datos$Age))
anova(a3)
```

```
## Analysis of Variance Table
##
## Response: as.integer(datos$Survived)
##                          Df  Sum Sq Mean Sq F value   Pr(>F)
## as.numeric(datos$Age)     1   1.936 1.93588  8.2427 0.004189 **
## Residuals               889 208.791 0.23486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
a4 <- aov(as.integer(datos$Survived) ~ as.integer(datos$Embarked))
anova(a4)
```

```
## Analysis of Variance Table
##
## Response: as.integer(datos$Survived)
##                             Df  Sum Sq Mean Sq F value    Pr(>F)
## as.integer(datos$Embarked)   1   5.925  5.9246  25.717 4.811e-07 ***
## Residuals                  889 204.803  0.2304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Encontramos un efecto significativo de, al menos, los campos `Pclass`,`Sex` y `Embarked`. El p-valor es, en los tres, menor que 0.05, por lo que se rechazan las hipótesis nulas y se puede asegurar que las medias son diferentes. Esto se interpreta de modo que, en comparación: * Sobrevivieron más mujeres que hombres. * Sobrevivieron más pasajeros de primera clase que de segunda y tercera. * La probabilidad de sobrevivir depende del puerto en el que se ha embarcado.

Por otra parte, el p-valor obtenido con respecto a la edad (`Age`), permitiría rechazar la hipótesis nula pero no de una manera tan clara. Estamos hablando de un orden de magnitud similar, por lo que no nos aventuramos a concluir que haya un efecto significativo de la edad. Esto significaría que el número medio de supervivientes fue similar para todas las edades.

¿Qué variables influyen más en la supervivencia de los pasajeros?

Nos planteamos, también, elaborar un modelo de regresión lineal con los factores anteriormente estudiados y algún otro. Para ello, consideramos crear dos grupos etarios: infantes y adultos. Tomando los 16 años como el punto de paso a la edad adulta (estamos hablando de principios del siglo XX). También incorporamos el dato de si el pasajero viaja solo o con familia, usando para ello los dos campos `SibSp` y `Parch`. Esto lo hacemos tras comprobar que dejando valores numéricos no aportan prácticamente nada al modelo de regresión.

Por tanto, ejecutamos todas estas operaciones y creamos el modelo con la función `lm`:

```r
datos$newAge[as.numeric(datos$Age) < 17] = "Children"
datos$newAge[as.numeric(datos$Age) >= 17] = "Adults"
datos$family[(datos$SibSp == "0" & datos$Parch == "0")] = "lonely"
datos$family[(datos$SibSp != "0" | datos$Parch != "0")] = "with family"
datos$Pclass <- as.factor(datos$Pclass)
datos$family <- as.factor(datos$family)
datos$newAge <- as.factor(datos$newAge)
modelo <- lm(Survived ~ Pclass + Sex + newAge + family + Embarked, data = datos)
```

```
## Warning in model.response(mf, "numeric"): using type = "numeric" with a
## factor response will be ignored
```

```
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors
```

```r
coefficients(modelo)
```

```
##       (Intercept)           Pclass2           Pclass3           Sexmale
##        1.97075786       -0.13151590       -0.32615642       -0.50051672
##    newAgeChildren familywith family         EmbarkedQ         EmbarkedS
##        0.14386176       -0.02338671       -0.01154101       -0.08829422
```

Algunas conclusiones que podríamos extraer de este modelo: viajar en segunda clase penaliza y en tercera penaliza aún más. Ser varón penaliza más que la clase, ser infante bonifica, viajar con familia es apenas irrelevante y los puertos de embarque también son poco "perjudiciales", aunque lo sean específicamente S y Q.

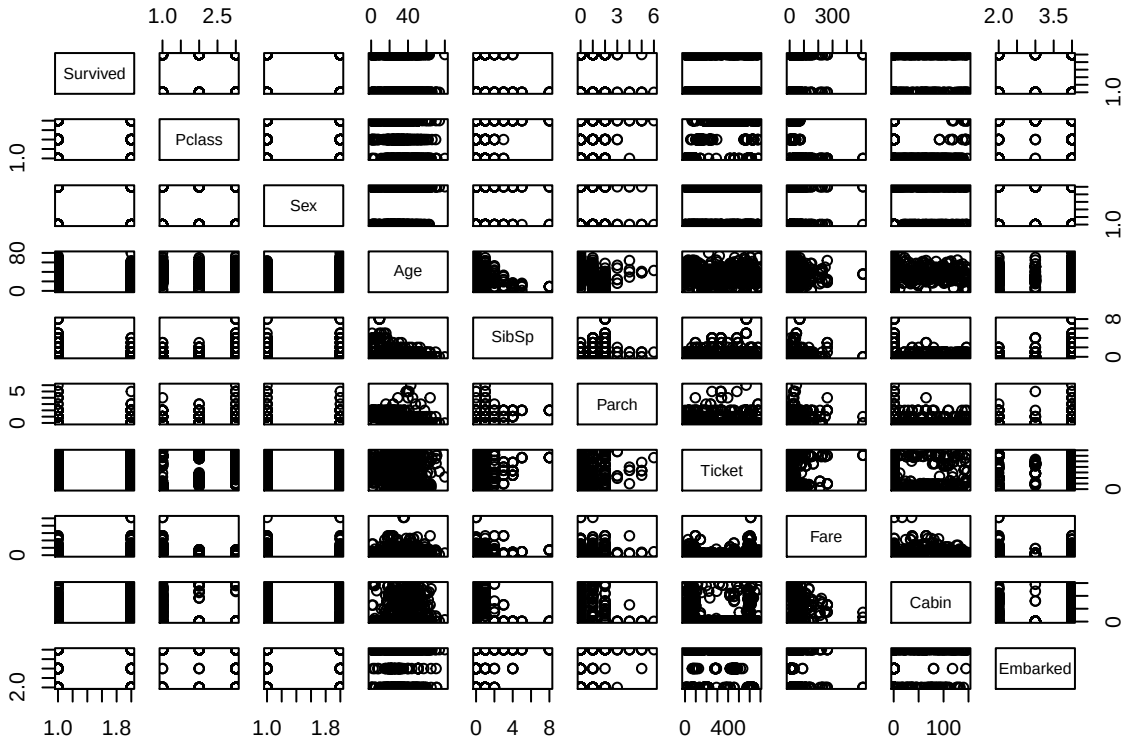
¿La probabilidad de sobrevivir es mayor dependiendo del sexo del pasajero? Es evidente que sí, que es mayor si eres mujer.

¿La probabilidad de sobrevivir es mayor viajando en primera clase? También.

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

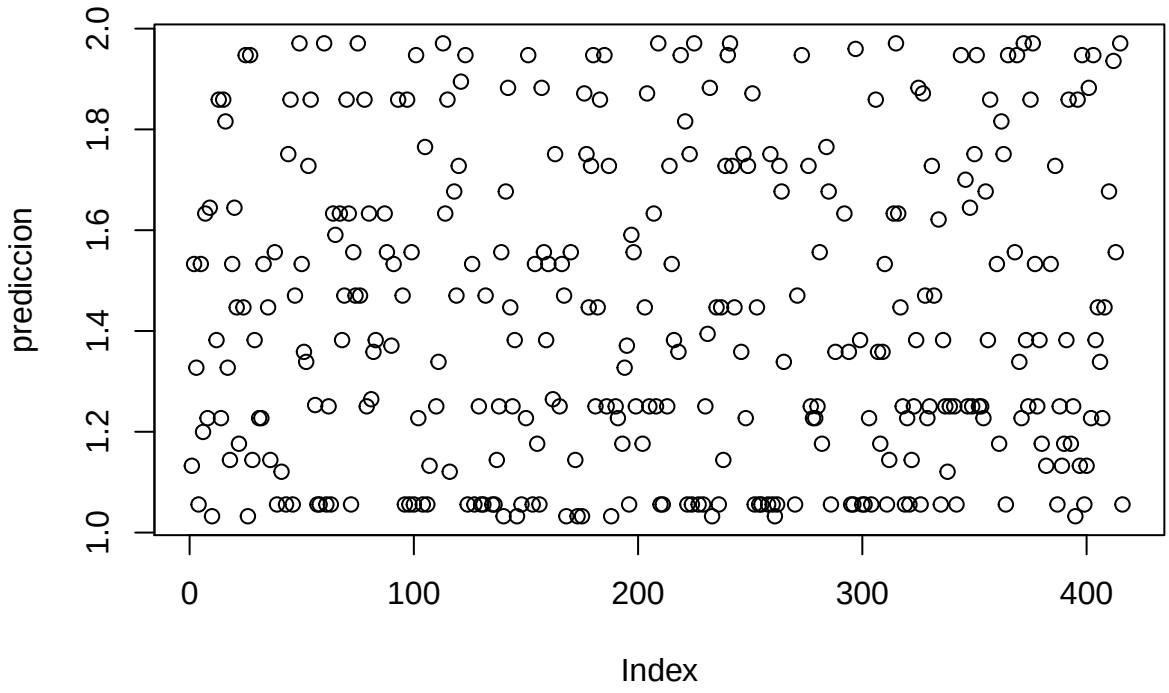
En primer lugar, podemos echar un vistazo a la distribución de los datos originales.

```r
plot(datos[,1:10])
```



Para, después, hacer una comprobación de si el modelo "predice" lo que intenta predecir o no. Para ello usamos el segundo set de datos proporcionado, el denominado `test`.

```
datosNuevos <- read.csv("all/test.csv", header = TRUE)
datosNuevos$newAge[as.numeric(datosNuevos$Age) < 17] = "Children"
datosNuevos$newAge[as.numeric(datosNuevos$Age) >= 17] = "Adults"
datosNuevos$family[(datosNuevos$SibSp == "0" & datosNuevos$Parch == "0")] = "lonely"
datosNuevos$family[(datosNuevos$SibSp != "0" | datosNuevos$Parch != "0")] = "with family"
datosNuevos$Pclass <- as.factor(datosNuevos$Pclass)
datosNuevos$family <- as.factor(datosNuevos$family)
datosNuevos$newAge <- as.factor(datosNuevos$newAge)
prediccion <- predict(modelo, datosNuevos)
plot(prediccion)
```




```
summary(prediccion)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.032   1.176   1.382   1.426   1.677   1.971      86
```

No hemos logrado averiguar por qué la predicción nos la sitúa entre 1 y 2, en lugar de entre 0 y 1. Suponemos que es por algún offset, pero no estamos seguros. Se pueden comprobar casos al azar para darse cuenta de que funciona. Por ejemplo, el registro 13 de este conjunto de datos:

```
datosNuevos[13,]
```

```
##    PassengerId Pclass                                          Name    Sex
## 13         904      1 Snyder, Mrs. John Pillsbury (Nelle Stevenson) female
##    Age SibSp Parch Ticket    Fare Cabin Embarked newAge      family
## 13  23     1     0  21228 82.2667   B45        S Adults with family
```

Que es mujer y viaja en primera clase. Nuestro modelo arroja un valor de 1,859 para su supervivencia (que interpretamos como 0.859).

```
prediccion[13]
```

```
##       13
## 1.859077
```

De esta forma, podríamos predecir cualquier caso.

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

Los estudios realizados permiten extraer conclusiones sobre la probabilidad de supervivencia de un pasajero en función de las variables conocidas.

Durante el proceso, se han tenido que imputar datos faltantes y comprobar la existencia de valores atípicos.

Se han realizado pruebas estadísticas sobre un conjunto de datos que se correspondían con diferentes variables relativas al pasaje del naufragio del Titanic.

Las variables utilizadas no se seleccionan de manera definitiva a priori, sino que se han ido ajustando según la necesidad a lo largo del proceso de limpieza y análisis.

Así, mediante un modelo de regresión lineal, podemos hacer predicciones de supervivencia o no según datos de pasajeros del set de datos.