Práctica 2 Limpieza y resolución de datos

Jone Aliri Lazcano, Nadia Nathaly Sánchez Pozo

23/12/2019

Índice

1. DETALLES DE LA ACTIVIDAD

1.1. Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2. Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

1.3 Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

2. RESOLUCIÓN

2.1. Descripción del dataset

La base de datos que se ha analizado en esta práctica se título **Titanic: Machine Learning from Disaster** (https://www.kaggle.com/c/titanic) [!https://www.kaggle.com/c/titanic].

Los datos de este archivo se encuentran divididos en dos bases de datos, uno de ellos es la base de datos de entrenamiento (*train.csv*) y la otra base de datos es la de prueba (*test.csv*), para este caso de estudio vamos a juntar los conjuntos de datos.

A partir de esta base de datos se plantea la problemática de determinar qué variables son más influyentes en la probabilidad de sobrevivencia de los pasajeros, mediante el estudio individual y colectivo de las variables, aplicando pruebas estadisticas.

Resumiendo, el 14 de Abril de 1992 el Titanic chocó con un iceberg y se llevo aproximadamente a 1500 de sus pasajeros y tripulación a las profundidades del oceano. Este incidente ha sido considerado uno de los desastres marinos más importantes en tiempos de paz, y a causa de dicho indicente se actualizaron o renovaron numerosas políticas de seguridad. Sin embargo, existen numerosas voces que dicen que hubo circunstancias que hicieron que hubiera un desproporcionada cantidad de muertos. El objetivo de analizar esta base de datos es explorar los factores que tuvieron relación con el hecho de que una persona sobreviviera o no a la catastrofe del Titanic.

Descríbamos la base de datos.

Esta base de datos tiene 891 observaciones y 12 variables.

- **PassengerId**:Variable que aporta el código de identificación de los pasajeros.
- **Survived**: Variable dicotómica que indica si el pasajero sobrevivión (1) o no sobrevivió (0).
- **Pclass**: Variable categórica que indica si los pasajeros tenían tickets de primera clase (1), segunda clase (2) o tercera clase (3). Obviamente los tickets más caros eran los de primera clase, seguidos de los de segunda clase y finalmente los de tercera clase.
- **Name**: Variable de tipo cadena con el nombre y apellidos de los pasajeros.
- **Sex**: Variable dicotómica que indica si el pasajero era un hombre (1) o una mujer (2).
- **Age**: Variable numérica que indica la edad de los pasajeros en años. En el caso de ser personas con menos de un años se indica la fracción (con un décimal), en caso de tener más de un año se utilizan números enteros.
- **SibSp**: Variable numérica (números enteros) que indicaba el número de familiares/cónyuges que tenían los pasajeros a bordo del Titanic.
- **Parch**: Variable numérica (números enteros) que indicaba el número de hijos/padres que tenían los pasajeros a bordo del Titanic.
- **Ticket**: Código/Número del tiket (podía haber más de un pasajero con el mismo número).
- **Fare**: Variable numérica que indica el precio del pase del pasajero.
- **Cabin**: Código que identifica la cabina del pasajero.
- **Embarked**: Variable categórica con tres niveles que indica el puerto en el cual embarcaron (puerto "C", "Q", o "S") que indican C = Cherbourg, Q = Queenstown y S = Southampton.

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

En primer lugar integraremos las dos bases de datos que tenemos (*train* y *test*). Antes de hacer esto, añadiremos una variable en cada base de datos para que identifique si la observación pertenece a la base *train* o a la base *test*. esta nueva variable la denominaremos *source*.

```
# cargar bases de datos
train <- read.csv("train.csv",stringsAsFactors = FALSE)
test <- read.csv("test.csv",stringsAsFactors = FALSE)
```

```
# Añadimos la variable source
train$source <- "train"
test$source <- "test"
```

```
# Unimos los dos conjuntos de datos en uno solo
complete_df  <- bind_rows(train,test)
filas=dim(train)[1]
```

Veamos qué es lo que tenemos en la base de datos completa. Para tener una idea general de los datos podemos utilizar las funciones **str()**, **summary()** y **dim()**.

```
dim(complete_df)
```

```
## [1] 1309   13
```

```
summary(complete_df)
```

```
##   PassengerId      Survived          Pclass          Name
##  Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:1309
##  1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median : 655   Median :0.0000   Median :3.000   Mode  :character
##  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
##      Sex                 Age            SibSp            Parch
##  Length:1309        Min.   : 0.17   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000
##  Mode  :character   Median :28.00   Median :0.0000   Median :0.000
##                     Mean   :29.88   Mean   :0.4989   Mean   :0.385
##                     3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##                     Max.   :80.00   Max.   :8.0000   Max.   :9.000
##                     NA's   :263
##     Ticket               Fare            Cabin             Embarked
##  Length:1309        Min.   :  0.000   Length:1309        Length:1309
##  Class :character   1st Qu.:  7.896   Class :character   Class :character
##  Mode  :character   Median : 14.454   Mode  :character   Mode  :character
##                     Mean   : 33.295
##                     3rd Qu.: 31.275
##                     Max.   :512.329
##                     NA's   :1
##     source
```

```
##  Length:1309
##  Class :character
##  Mode  :character
##
##
##
##
```

```
str(complete_df)
```

```
## 'data.frame':    1309 obs. of  13 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       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ 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     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
##  $ source     : chr  "train" "train" "train" "train" ...
```

Bien, vemos que el conjunto de datos completo contiene 1309 registros y 13 variables

2.3. Limpieza de los datos.

Una de las cosas importantes a la hora de importar los datos es ver si R ha asignado correctamente la categoría a cada varaiable. Por ejemplo, si nos fijamos en la variable *Survived* podemos ver que para R es una varaiable de tipo integer (numérico), pero quizás sería mejor que fuera un factor con dos niveles, donde 0 indique que no sobrevivió y 1 indique que sí sobrevivió. Así mismo, la variable *Pclass* también sería un factor, en este caso un factor ordenado ya que es una variable ordinal.

Hagamos estos cambios para luego poder trabajar mejor.

```
complete_df <- complete_df %>%
  mutate(Survived = as.factor(Survived),
         Pclass = as.factor(Pclass),
         SibSp = as.factor(SibSp),
         Parch = as.factor(Parch))
```

2.3.1. Ceros o elementos vacíos.

```
# Estadísticas de valores vacíos
colSums(is.na(complete_df))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0         418           0           0           0         263
```

```
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
##      source
##           0
```

```
colSums(complete_df=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0          NA           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
##      source
##           0
```

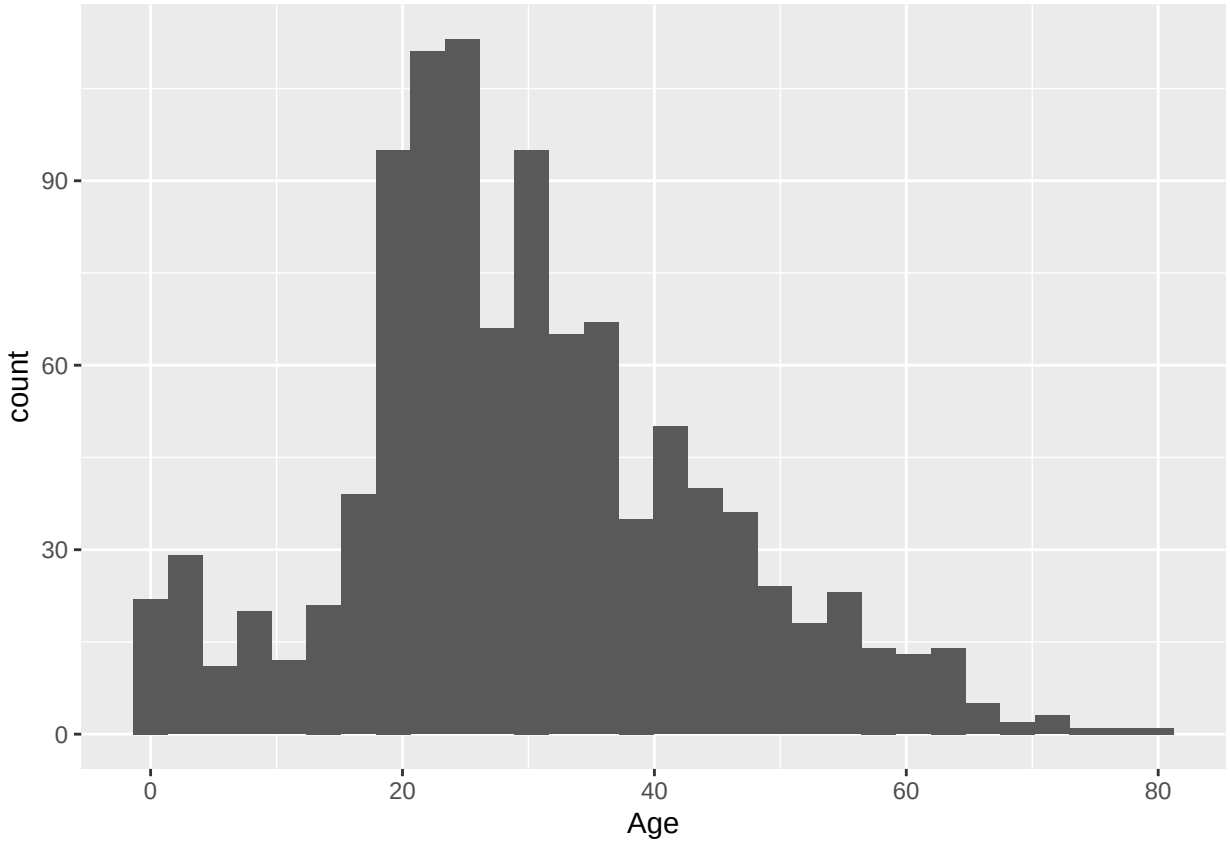
Tenemos cinco variables con datos perdidos o elementos vacios, las variables: *Survived*, *Age*, *Fare*, *Cabin* y *Embarked*. La variable *Survived* contiene valores perdidos debido a que desconocemos dicho datos de todas las observaciones de la base de datos *test*; La variable *Age* es una variable cuantitativa donde tenemos 263 valores perdidos que están indicados con "NA"; La variable *Fare* es una variable cuantitativa y tiene un valor perdido que esta indicado como NA (esta variable indica el precio del ticket, y no sabemos por qué hay bastantes sujetos que tienen un ticket que costó 0 dolares. Este dato es raro, puede que sea una errata y que quiera indicar un valor perdido, o pueda ser que realmente no se pago ese ticket debido a que son personas de la tripulación, o sean niños, u otras razones que por ahora desconocemos. Quizás haya que investigar más a fondo este dato); finalmente, las variables *Cabin* y *Embarked*, están caracterizados como factores y tienen un nivel que indica un valor perdido. Así, en la variable *Cabin* tenemos 1014 observaciones con valores perdidos, y en la variable *Embarked* 2 observaciones con valores perdidos. Las variables *Survived*, *SibSp* y *Parch* también contienen ceros, pero esos valores no son valores perdidos y tienen un sentido claro (que no sobrevivió, que no tenía hermanos/as ni esposo/a a bordo, y que no tenía padres ni hijos/as a bordo).

Empecemos con la variable *Age*. Esta variable es una variable cuantitativa e indica la edad de los pasajeros. Empecemos por explorar la distribución de esta variable en nuestra muestra.

```
complete_df %>%
  ggplot(aes(Age)) +
  geom_histogram()
```

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

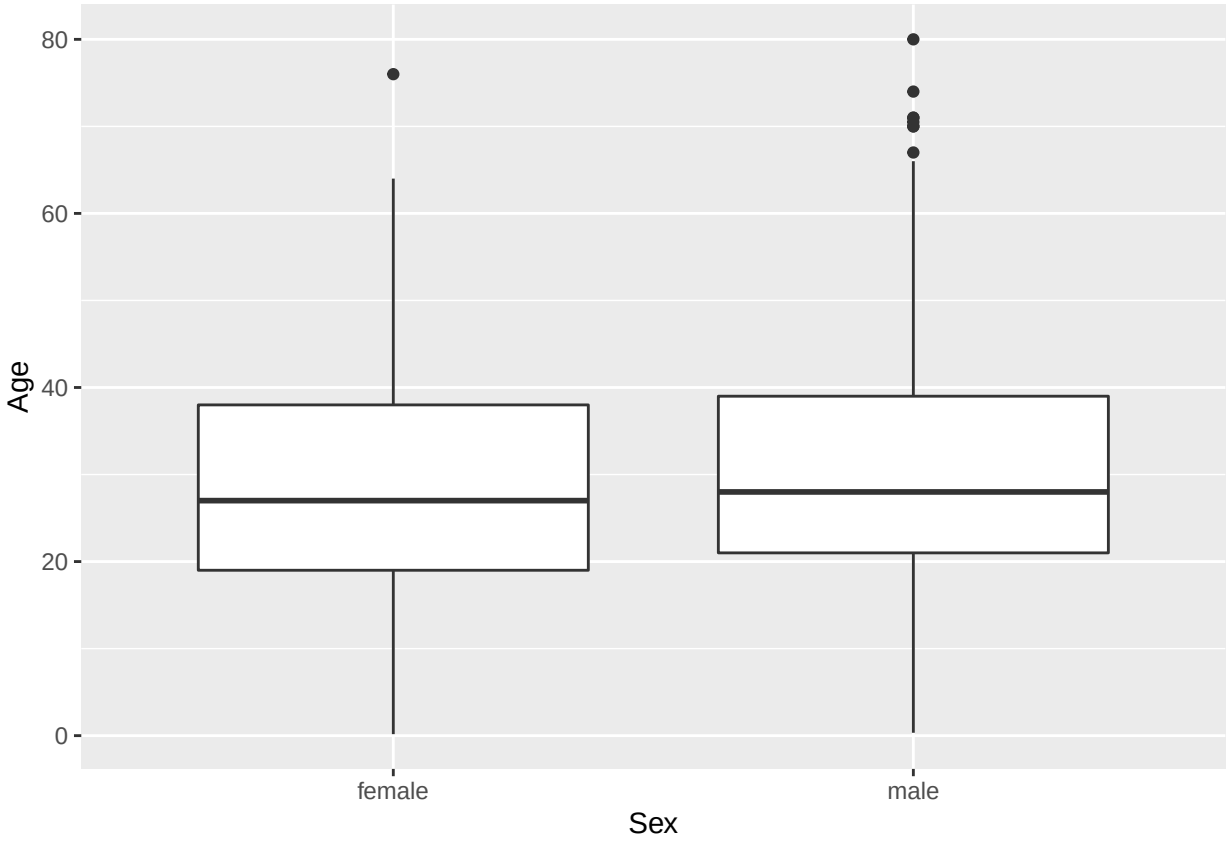
```
## Warning: Removed 263 rows containing non-finite values (stat_bin).
```



Bien, podemos observar que la mayoría de los sujetos que tenemos se encuentran entre los 20 y los 40 años. Una de las opciones de imputar los datos de esta variable sería utilizar la media del grupo. Pero antes de hacer esto, veamos si por ejemplo la distribución es muy diferente en función del género, ya que en ese caso en cada género podríamos imputar la media de su grupo.

```
complete_df %>%
  ggplot(aes(Sex, Age)) +
  geom_boxplot()
```

```
## Warning: Removed 263 rows containing non-finite values (stat_boxplot).
```



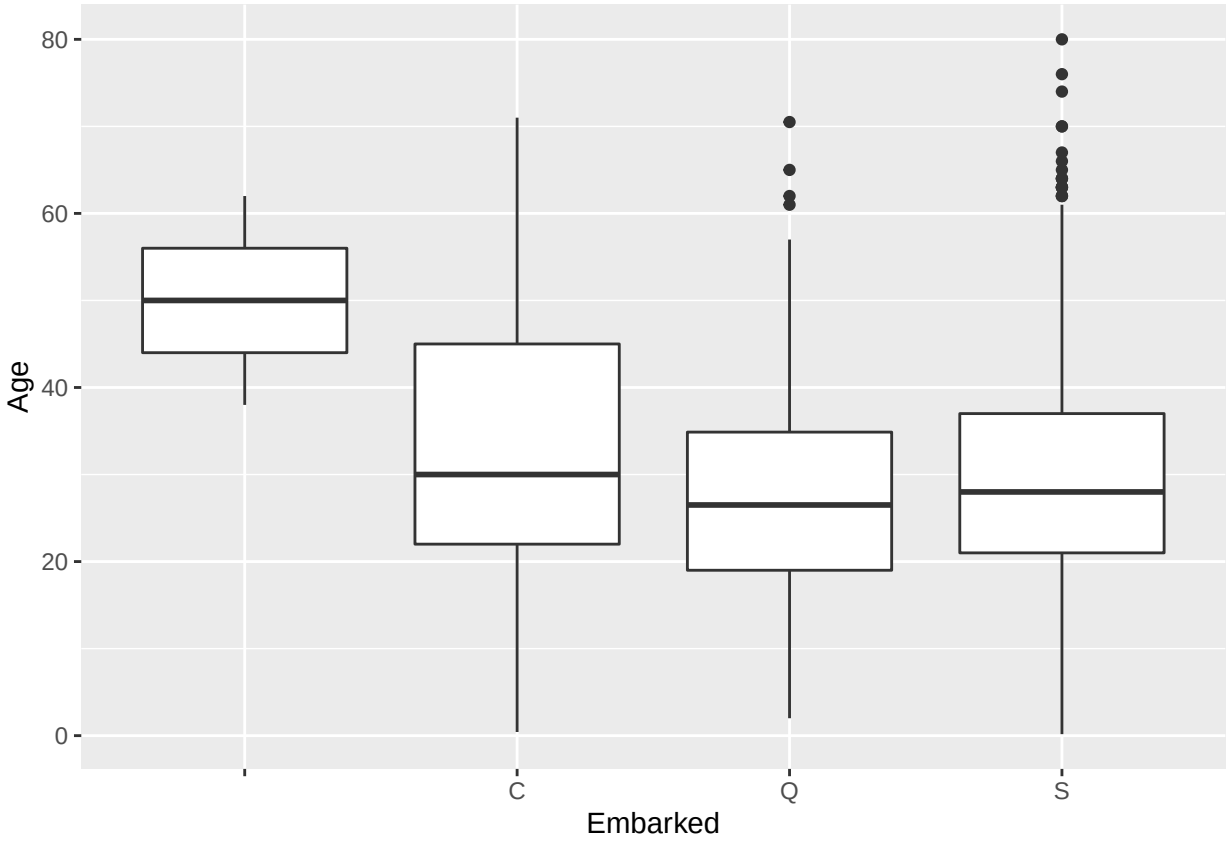

```
table(complete_df$Sex)
```

```
##
## female   male
##    466    843
```

Podemos observar que la distribución y es realmente parecida en ambos géneros, por lo que no mejoraría la imputación el hecho de utilizar las medias de cada grupo de género. Podemos analizar por ejemplo la distribución de esta variable en función del puerto de embarque:

```
complete_df %>%
  ggplot(aes(Embarked, Age)) +
  geom_boxplot()
```

```
## Warning: Removed 263 rows containing non-finite values (stat_boxplot).
```



Sin embargo, sí que vemos algunas diferencias en función del puerto de embarque. Sobre todo en el caso de las observaciones que tienen un valor perdido en el puerto de embarque. Sin embargo, antes de confiar en este gráfico analicemos cuántas personas forman cada grupo, ya que si el número de personas es reducido, el utilizar esa media no sería la mejor opción (ya que dicho dato estaría sobredimensionado y no sería representativo).

```
table(complete_df$Embarked)
```

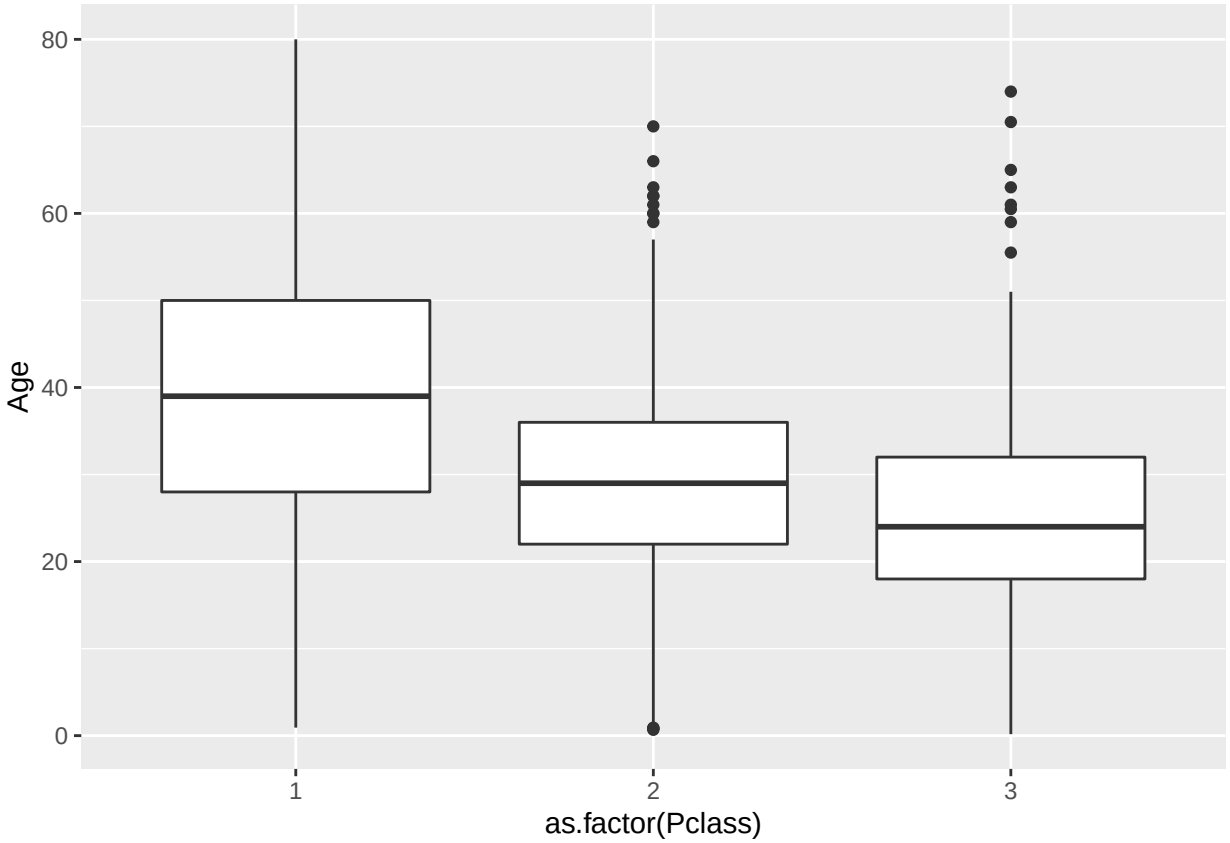
```
##
##       C   Q   S
##   2 270 123 914
```

Efectivamente, vemos que son tan solo dos observaciones las que conforman el "grupo" de las personas que no tienen dato sobre el puerto de embarque, por lo que la media de edad de este "grupo" no sería el ideal a la hora de realizar imputaciones. En cuanto a los demás grupos vemos que el puerto de embarque S es el que más observaciones tiene, y viendo que los otras dos puertos tienen relativamente pocas observaciones y las diferencias no son excesivas, no parece que mejoraríamos mucho la imputación de la edad utilizando las medias de cada uno de esos grupos.

Sigamos analizando la edad en función de otras variables como la clase.

```
complete_df %>%
  ggplot(aes(as.factor(Pclass), Age)) +
  geom_boxplot()
```

```
## Warning: Removed 263 rows containing non-finite values (stat_boxplot).
```




```
table(complete_df$Pclass)
```

```
##
##   1   2   3
## 323 277 709
```

Podemos observar que el rango de edad es algo más amplio, y existe mayor desviación en los pasajeros de primera clase, y que los pasajeros de segunda y tercera clase tienden a ser algo más jóvenes (aunque no sabemos si esas diferencias son significativas o no). Sin embargo de nuevo, el número de sujetos es bastante más grande en tercera clase (y ya se sabe que a mayor número de observaciones la tendencia es que la varianza sea más pequeña). Por ello, se vuelve a concluir que imputar las medias de edad en función de la clase no mejoraría enormemente la calidad de la imputación.

Podríamos analizar la distribución de la edad en función de las variables *Parch* y *SibSp*, pero éstas tienen unas distribuciones muy desiguales, y las medias obtenidas de grupos no muy numerosos no son estables y por lo tanto no son útiles para realizar imputaciones.

```
table(complete_df$Parch)
```

```
##
##    0    1    2    3    4    5    6    9
## 1002  170  113    8    6    6    2    2
```

```
table(complete_df$SibSp)
```

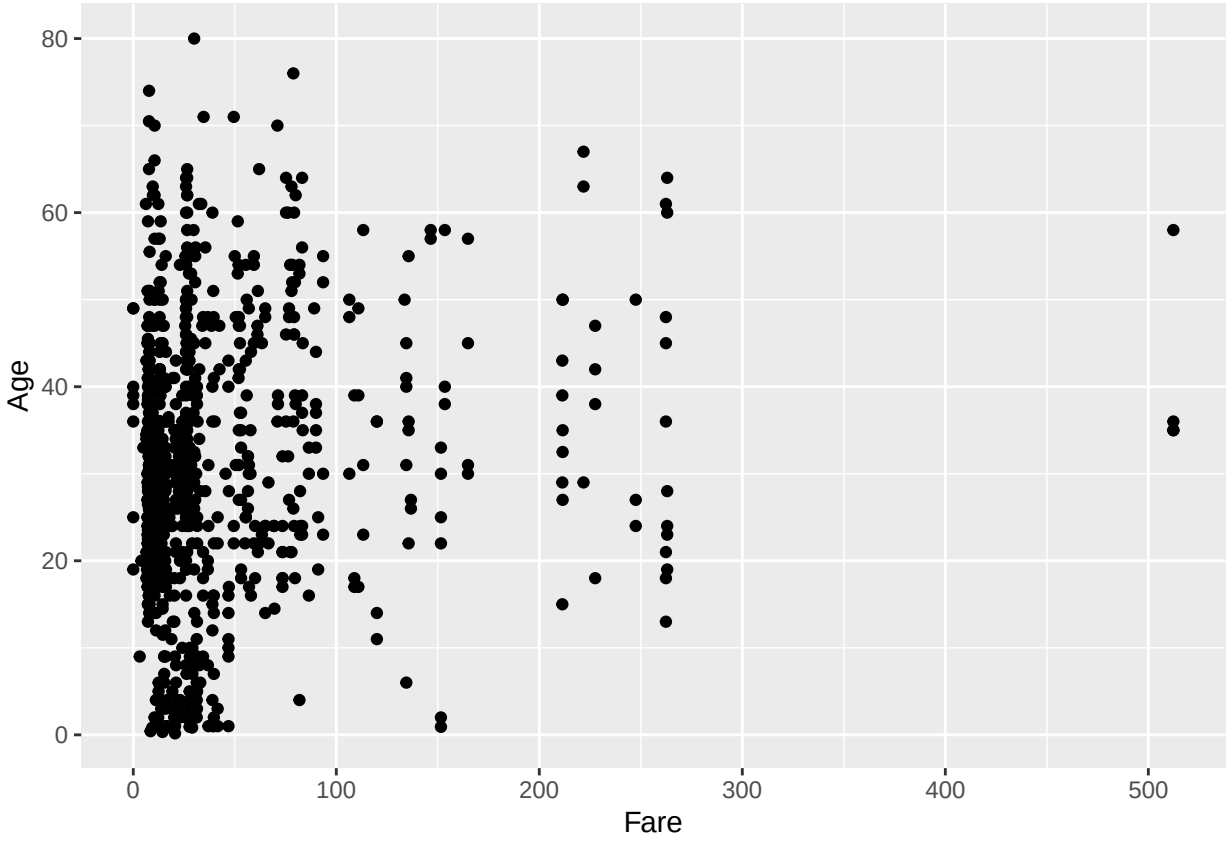
```
##
##   0   1   2   3   4   5   8
## 891 319  42  20  22   6   9
```

Sí, tal y como podemos apreciar en ambos casos, las frecuencias son muy elevadas en una categoría, y bastante más reducidas en las demás, por lo que no realizaremos imputaciones en función de estas variables.

Por último analizaremos la relación de la edad con la tarifa, por si existe una alta relación y podemos utilizar ese dato para realizar una imputación (por ejemplo utilizando una regresión donde la tarifa como variable predictora sea capaz de predecir la edad).

```
complete_df %>%
  ggplot(aes(Fare, Age)) +
  geom_point()
```

```
## Warning: Removed 264 rows containing missing values (geom_point).
```




```
cor.test(complete_df$Fare, complete_df$Age)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  complete_df$Fare and complete_df$Age
## t = 5.867, df = 1043, p-value = 5.955e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1193909 0.2368160
## sample estimates:
##       cor
## 0.1787399
```

Tal y como se concluye del gráfico y del resultado de la correlación vemos que estas dos variables no están relacionadas y por lo tanto no podemos utilizar la tarifa para predecir la edad.

Por lo realizado hasta ahora, se concluye que imputar los valores perdidos en la variable edad utilizando una única media (la media correspondiente al grupo entero) sería la mejor solución por lo tanto eso es lo que haremos. Para realizar las imputaciones utilizaremos la función **mutate()** e indicaremos que en la variable Age en todos aquellos casos en los que el valor sea un valor perdido, se reemplace este valor por la media de la variable Age, también tenemos que indicar que a la hora de calcular la media descarte los valores perdidos ya que de lo contrario imputaríamos valores perdidos.

```
complete2_df <- complete_df %>%
  mutate(Age = ifelse(is.na(Age), mean(complete_df$Age, na.rm=TRUE), Age))
```

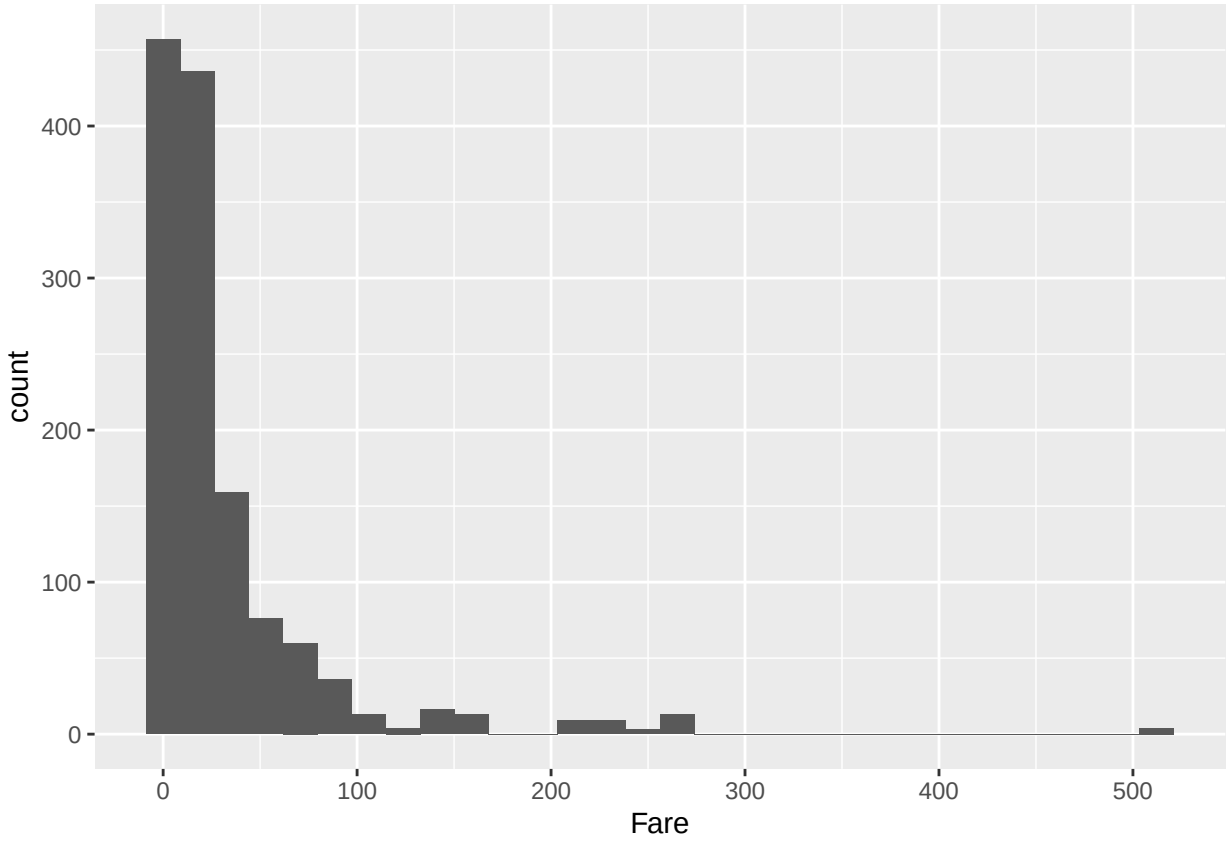
Sigamos analizando los valores perdidos y realizando las imputaciones. La siguiente variable que vamos a analizar es la variable *Fare*. Tal y como hemos comentado anteriormente, esta variable tiene un valor perdido que está indicado como NA, pero también tiene otras observaciones con un valor de 0. Hemos comentado que esto en principio es extraño, y que podria ser que los tripulantes por ejemplo tengan ese valor. En el caso del valor perdido, una imputación viable sería utilizar la media de todo el grupo, pero en el caso de que se concluyera que las observaciones que tienen 0 son tripulantes (y no pasajeros), lo lógico sería calcular la media sin esos valores en caso de que pensáramos que la observación que tiene el valor perdido sea pasajero, o imputar directamente un 0 en caso de que pensáramos que la observación que tiene el valor perdido es un tripulante.

Veamos en primer lugar la distribución de la variable:

```
complete2_df %>%
  ggplot(aes(Fare)) +
  geom_histogram()
```

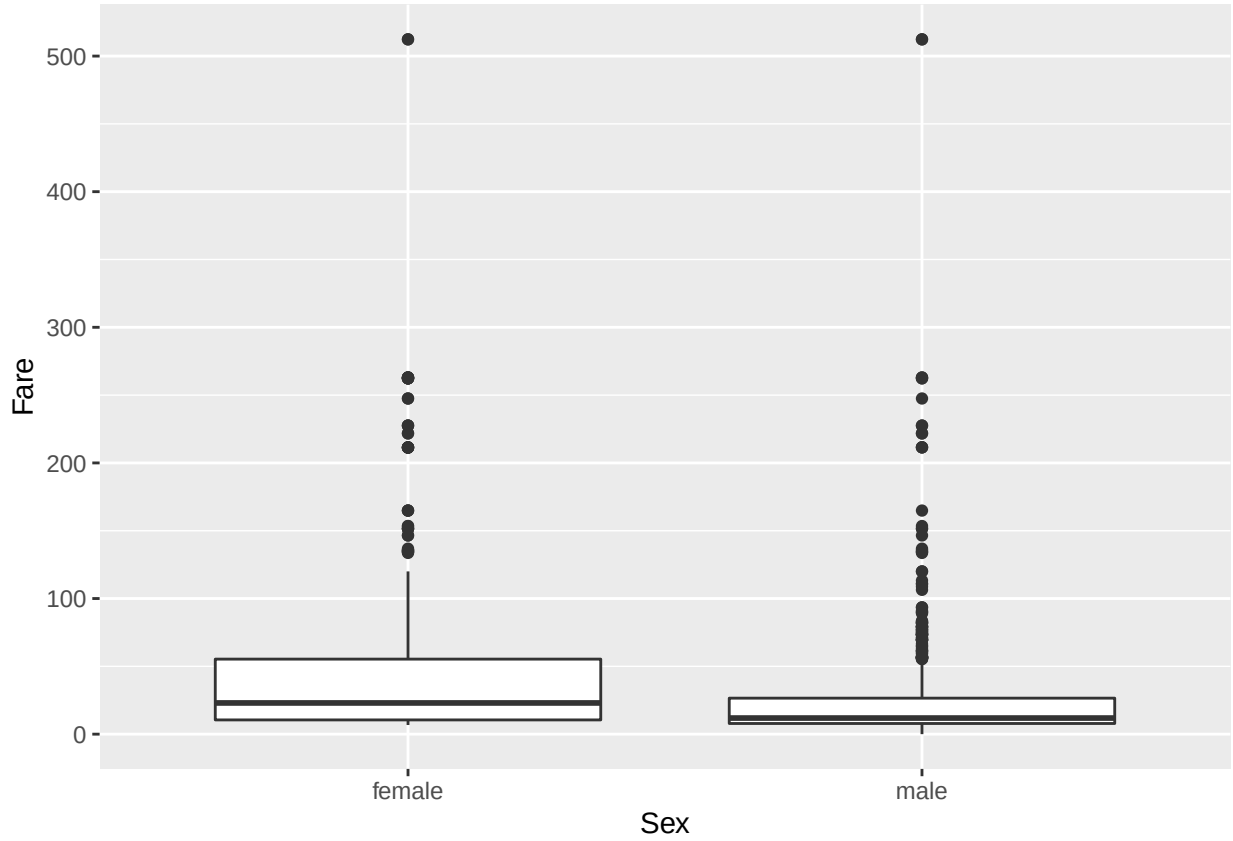
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```




```
complete2_df %>%
  ggplot(aes(Sex, Fare)) +
  geom_boxplot()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



Podemos observar que hay algunos valores extremos que tendríamos que tener en cuenta a la hora de calcular la media para realizar la imputación.

Analicemos a los sujetos que tienen el valor 0 en esta variable.

```
gratis <- complete2_df[complete2_df$Fare == 0,]
```

```
summary(gratis)
```

```
##   PassengerId     Survived   Pclass      Name              Sex
##  Min.   : 180.0   0   :14   1   :7   Length:18          Length:18
##  1st Qu.: 303.0   1   : 1   2   :6   Class :character   Class :character
##  Median : 598.0   NA's: 3   3   :4   Mode  :character   Mode  :character
##  Mean   : 598.1             NA's:1
##  3rd Qu.: 807.0
##  Max.   :1264.0
##  NA's   :1
##       Age            SibSp        Parch       Ticket              Fare
##  Min.   :19.00   0      :17   0      :17   Length:18          Min.   :0
##  1st Qu.:29.88   1      : 0   1      : 0   Class :character   1st Qu.:0
##  Median :29.88   2      : 0   2      : 0   Mode  :character   Median :0
##  Mean   :33.17   3      : 0   3      : 0                      Mean   :0
##  3rd Qu.:38.00   4      : 0   4      : 0                      3rd Qu.:0
##  Max.   :49.00   (Other): 0   (Other): 0                      Max.   :0
##  NA's   :1       NA's   : 1   NA's   : 1                      NA's   :1
##     Cabin             Embarked            source
```

```
##  Length:18          Length:18          Length:18
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

Bien por los descriptivos obtenidos, podemos decir que la mayoría de los pasajeros que tienen el valor 0 en esta la variable *Fare* no sobrevivieron, eran varones, mayores de edad (entre 19 y 49 con una media de 33), no tenian ni esposas ni hijos/as a bordo y embarcaron en Southampton. Por esos datos podría ser que fueran miembros de la tripulación, o miembros de la banda de música o algún otro gremio que fuera en el barco pero no fuera pasajero como tal.

Por todo ello, vamos a calcular la media de la variable sin tener en cuenta los valores muy extremos. Una opción sería utilizar la media recortada al 5 % o al 10 %. Calculemos la media, y las dos medias recortadas que acabamos de mencionar.

```
#media
mean(complete2_df$Fare, na.rm = TRUE)
```

```
## [1] 33.29548
```

```
#la media recortada al 10%
mean(complete2_df$Fare,trim=10/100, na.rm = TRUE)
```

```
## [1] 21.57439
```

```
#la media recortada al 5%
mean(complete_df$Fare,trim=5/100, na.rm = TRUE)
```

```
## [1] 24.73131
```

Vista las diferencias optaremos por utilizar la media recortada al 5 % para imputar en el único valor perdido que teníamos en esta variable.

```
complete2_df <- complete2_df %>%
  mutate(Fare = ifelse(is.na(Fare), mean(complete2_df$Fare, trim = 5/100, na.rm=TRUE), Fare))
```

Seguimos analizando los valores perdidos, ahora nos toca la variable *Cabin*.

```
camarote <- complete2_df[complete2_df$Cabin == "",]
dim(camarote)
```

```
## [1] 1014   13
```

Podemos ver que tenemos 1014 sujetos sin el dato del camarote. Realmente este dato es un dato categórico que no tiene mucho sentido imputar, ya que no tiene lógica que pongamos que unos sujetos estaban en un camarote X utilizando para ello el dato de en qué camarote estaban otros sujetos. Por lo tanto no vamos a imputar nada en estos casos. Sin embargo, en la variable *Embarked* que también es un factor sí que vamos a imputar un valor. Concretamente la variable *Embarked* indica el puerto en el que se han embarcado y puede tener cuatro valores. Por lo tanto calcularemos una tabla de frecuencias y veremos si hay un puerto que tenga una mayor frecuencia, en cuyo caso imputaremos dicho valor a los perdidos.

```
table(complete2_df$Embarked)
```

```
##
##       C   Q   S
##   2 270 123 914
```

Podemos observar que la mayoría de las personas embarcaron en Southampton por lo tanto imputaríamos S a las dos observaciones que tienen los valores perdidos.

```
complete2_df$Embarked[complete2_df$Embarked == ""] <- "S"
```

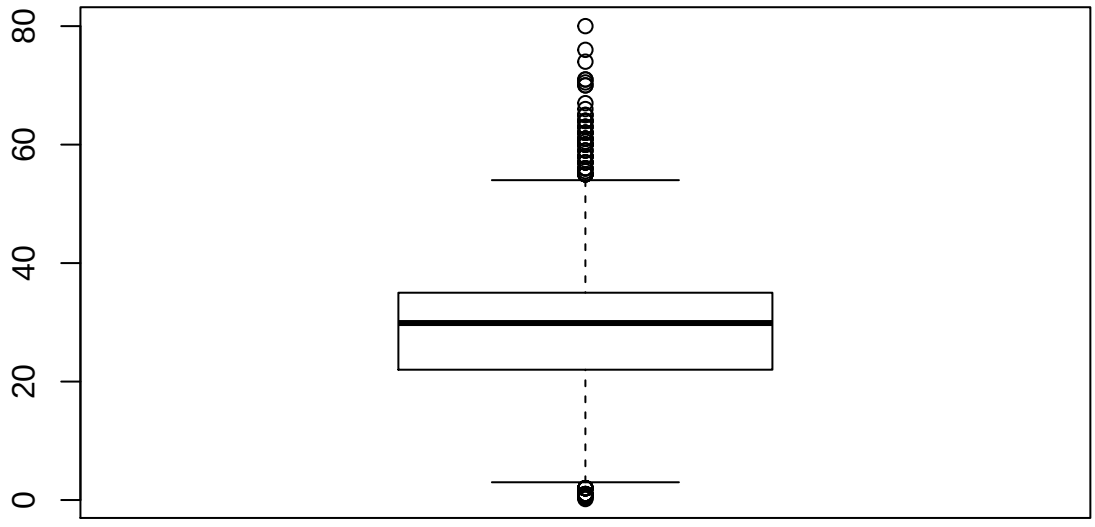
Finalmente la variable *Survived*. Esta variable tiene valores perdidos debido a que se han juntado los archivos train y test y las observaciones del archivo test no tienen este valor.

Sin embargo, todavía no vamos a imputar los valores de esta variable. Primero vamos a analizar mejor la base de datos y al final crearemos un modelo para realizar esta imputación.

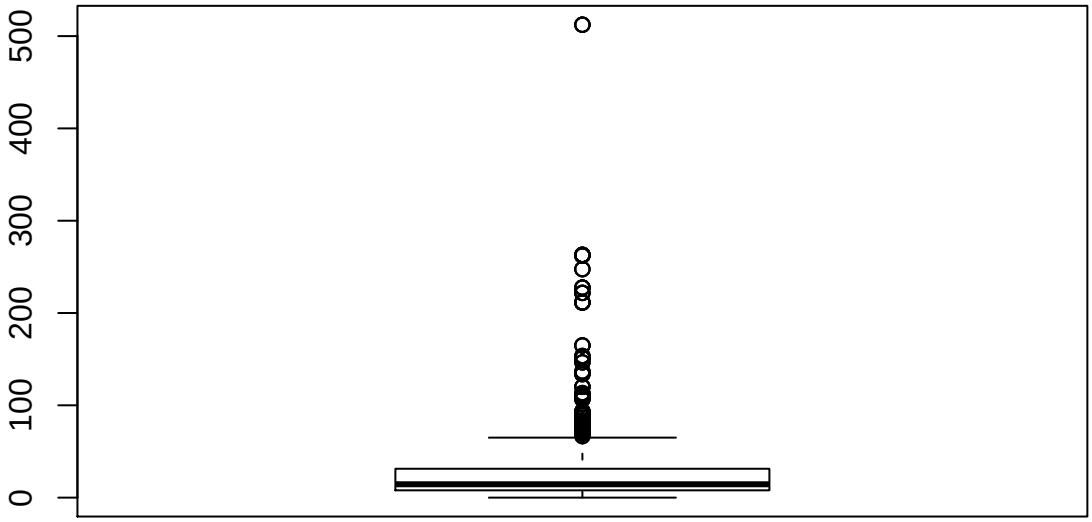
2.3.2. Identificación y tratamiento de valores extremos.

Tal y como hemos observado antes, tenemos valores perdidos en la variable *Fare* y en la variable *Age*.

```
boxplot(complete2_df$Age)
```



```r
boxplot(complete2_df$Fare)
```



Aunque son valores extremos, debido a que se alejan de la media y de los valores del resto del grupo, son valores que son posibles (están dentro del rango de posibles valores en esas variables), por lo tanto, a priori no los vamos a eliminar. Sin embargo tendremos que ver en los análisis posteriores si podemos seguir usándolos o debemos prescindir de ellos en algún caso.

2.3.3. Exportación de los datos preprocesados

Luego de realizar los procedimientos de integración, validación y limpieza, procedemos a guardar nuestros datos en un nuevo fichero denominado *titanic_clean.csv*

```r
# Exportación de los datos limpios en .csv

write.csv(complete2_df, "titanic_clean.csv")
```

2.4. Análisis de los datos.

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

En este caso práctico, trabajaremos con cuatro variables de entrada y una variable de respuesta. Las variables de entrada son: Age, PClass, Sex y Fare. La variable de respuesta es si sobrevivieron o no.

Seleccionamos el conjunto de datos de interes

```
df <- filter( complete2_df, complete2_df$source=="train")
df = df[,c(2,3,5,6,10)]
str(df)
```

```
## 'data.frame':    891 obs. of  5 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     : chr  "male" "female" "female" "female" ...
##  $ Age     : num  22 38 26 35 35 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

También añadimos una nueva variable Age1, la cual será categórica y convertimos la variable Sex a factor.

- Si age <= 18, entonces age = joven
- Si 18 < age <= 64, entonces age = adult
- Si age > 64, entonces age = senior

```
df$Age1 <- df$Age
df$Age1[df$Age1 <= 18] = "joven"
df$Age1[(df$Age1 > 18)& (df$Age1 <=64) ] = "adult"
df$Age1[(df$Age1 != "joven") & (df$Age1 != "adult")] = "senior"
df$Age1 = as.factor(df$Age1)
df$Sex = as.factor(df$Sex)

summary(df)
```

```
##  Survived Pclass      Sex           Age             Fare            Age1
##  0:549    1:216   female:314   Min.   : 0.42   Min.   :  0.00   adult :741
##  1:342    2:184   male  :577   1st Qu.:22.00   1st Qu.:  7.91   joven :139
##           3:491                Median :29.88   Median : 14.45   senior: 11
##                                Mean   :29.74   Mean   : 32.20
##                                3rd Qu.:35.00   3rd Qu.: 31.00
##                                Max.   :80.00   Max.   :512.33
```

A continuación, se especifican los ánalisis que podemos realizar en torno a nuestro conjunto de datos:

Ánalisis 1: comparar cuántos han sobrevivido en función del género y en función de la clase (primera, segunda o tercera) en la que estaban. Este análisis requeriría realizar una tabla de contingencia y una chi cuadrado, para ver si ambas variables están relacionadas. Este análisis es no parámetrico por lo tanto no habría que comprobar los supuestos de normalidad y homogeneidad de la varianza para realizar este análisis.

Ánalisis 2: Analizar si existen diferencias en la edad entre los sujetos que han sobrevivido y los que no han sobrevivido. Para ello realizaríamos una prueba de diferencia de medias. Podríamos hacer una prueba t (si se cumplen los supuestos de homogeneidad de varianzas y de normalidad de la variable dependiente), o si no se cumplen los supuestos podríamos hacer la prueba U de Mann-Whitney.

Ánalisis 3: Calcular un modelo de regresión regresión logística para predecir la supervivencia dadas las variables explicativas, edad, género, precio del pasaje, clase en la que viajó.

2.4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Comprobemos la normalidad de la variable edad. Existen muchas opciones de comprobar la normalidad, una opción viable sería utilizar la prueba de Kolmogorov-Smirnov con la corrección de Lilliefors.

```
lillie.test(df$Age)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Age
## D = 0.14957, p-value < 2.2e-16
```

```
lillie.test(df$Age[df$Survived == 0])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Age[df$Survived == 0]
## D = 0.18493, p-value < 2.2e-16
```

```
lillie.test(df$Age[df$Survived == 1])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  df$Age[df$Survived == 1]
## D = 0.096301, p-value = 4.268e-08
```

Los resultados, al igual que el diagrama de cajas, muestran que la distribución de la variable edad no es normal ni en la muestra en su conjunto (K-S(891)= 0,16; p < 0,001), ni en el subconjunto de los no supervivientes (KS(549)=0,18; p < 0,001) ni en el subconjunto de supervivientes (KS(342)=0,10; p < 0,001).

Comprobemos ahora la homogeneidad de las varianzas de la edad en los dos grupos que queremos comparar. Para ello podemos utilizar la prueba F de Levene.

```
leveneTest(df$Age, df$Survived)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  5.5007 0.01923 *
##       889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tal y como se puede ver en los resultados no se cumple el supuesto de homogeneidad de varianzas (F(1,889)=5,50; p=0,02).

A la vista de estos resultados tenemos dos opciones, utilizar la prueba U de Mann Whitney que como es una prueba paramétrica no necesita que se cumplan los supuestos analizados; o utilizar la prueba t con la indicando en el análisis que estamos utilizando varianzas no iguales. Es cierto que no se cumple la normalidad, pero teniendo una muestra tan amplia podríamos no tener en cuenta este hecho ya que por una parte la prueba t es robusta ante el inclumplimiento de este supuesto, sobre todo con muestras grandes.

2.4.3. Aplicación de pruebas estadísticas

2.4.3.1. ¿Cuántos han sobrevivido en función del género y en función de la clase?

Survived vs Sex Para contrastar la hipótesis de independencia entre género y sobrevivencia, usaremos la función chisq.test de R

```
summary(df$Sex)
```

```
## female   male
##    314    577
```

```
cuadro<-table(df$Survived, df$Sex)
chisq.test(cuadro)$expected
```

```
##
##       female     male
##   0 193.4747 355.5253
##   1 120.5253 221.4747
```

```
chisq.test(cuadro)
```

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

Al tener un p-valor muy reducido podemos deducir que el género y sobrevivencia no son independientes. Observamos que las sobrevivientes femeninas eran mucho más numerosas que los hombres.

Survived vs Pclass Para contrastar la hipótesis de independencia entre clase y sobrevivencia, usaremos la función chisq.test de R

```
summary(df$Pclass)
```

```
##   1   2   3
## 216 184 491
```

```
cuadro<-table(df$Survived, df$Pclass)
chisq.test(cuadro)$expected
```

```
##
##             1         2        3
##   0 133.09091 113.37374 302.5354
##   1  82.90909  70.62626 188.4646
```

```
chisq.test(cuadro)
```

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

Al tener un p-valor muy reducido podemos deducir que clase y sobrevivencia no son independientes.

Observamos que los pasajeros que los sobrevivientes que viajavan en primera y segunda clase tuvieron mayor probabilidad de supervivencia.

2.4.3.2.¿Existen diferencias significativas de edad entre los supervivientes?

Esta prueba consistirá en un contraste de hipótesis sobre dos muestras para determinar si existen diferencias de edad entre los sujetos que han sobrevivido y los que no.

A continuación se plantea el siguiente contraste de hipótesis de dos muestras sobre la diferencia de medias,

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

donde μ es igual a la media de edad de edad sujetos que han sobrevivido. μ_0 es igual a la media de edad de edad sujetos que no han sobrevivido.

Al tener una muestra de tamaño superior a 30, consideramos aplicar el t test.

```
t.test(complete2_df$Age[complete2_df$Survived == 0], complete2_df$Age[complete2_df$Survived == 1], pair
```

```
##
##  Welch Two Sample t-test
##
## data:  complete2_df$Age[complete2_df$Survived == 0] and complete2_df$Age[complete2_df$Survived == 1]
## t = 2.0534, df = 668.84, p-value = 0.04042
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.08228709 3.67589192
## sample estimates:
## mean of x mean of y
##  30.45654  28.57745
```

Puesto que obtenemos un p-valor menor a 0.05, con un nivel de confianza del 95 % podemos concluir que existen diferencias significativas de edad entre los sujetos que han sobrevivido y los que no han sobrevivido.

2.4.3.2. Modelo de regrsión regresión logística

se calculará un modelo de regresión regresión logística utilizando regresores tanto cuantitativos como cualitativos, para analizar los regresores con mayor influencia al realizar las predicciones de los supervivientes.

Los regresores cualitativos Edad (joven, adulto, senior), género, clase.

```
modelo <- glm(Survived  ~ Age1 +Sex + Pclass + Fare, data = df ,family=binomial())
summary(modelo)
```

```
##
## Call:
## glm(formula = Survived ~ Age1 + Sex + Pclass + Fare, family = binomial(),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4833  -0.7050  -0.4202   0.6879   2.2268
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.161913   0.279578   7.733 1.05e-14 ***
## Age1joven    0.699790   0.239449   2.923  0.00347 **
## Age1senior  -1.488888   1.077150  -1.382  0.16690
## Sexmale     -2.569994   0.186436 -13.785  < 2e-16 ***
## Pclass2     -0.871133   0.275041  -3.167  0.00154 **
## Pclass3     -1.983623   0.260075  -7.627 2.40e-14 ***
## Fare         0.001153   0.002082   0.554  0.57962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  814.72  on 884  degrees of freedom
## AIC: 828.72
##
## Number of Fisher Scoring iterations: 5
```

```
sel <- which(summary(modelo)$coefficients[-1,4] < 0.05)
sel <- sel + 1
```

Ha sido significativo el test parcial sobre el coeficiente de Age1joven, Sexmale, Pclass2, Pclass3. Siendo la estimacion de su coeficiente 0.7, -2.57, -0.871, -1.984.

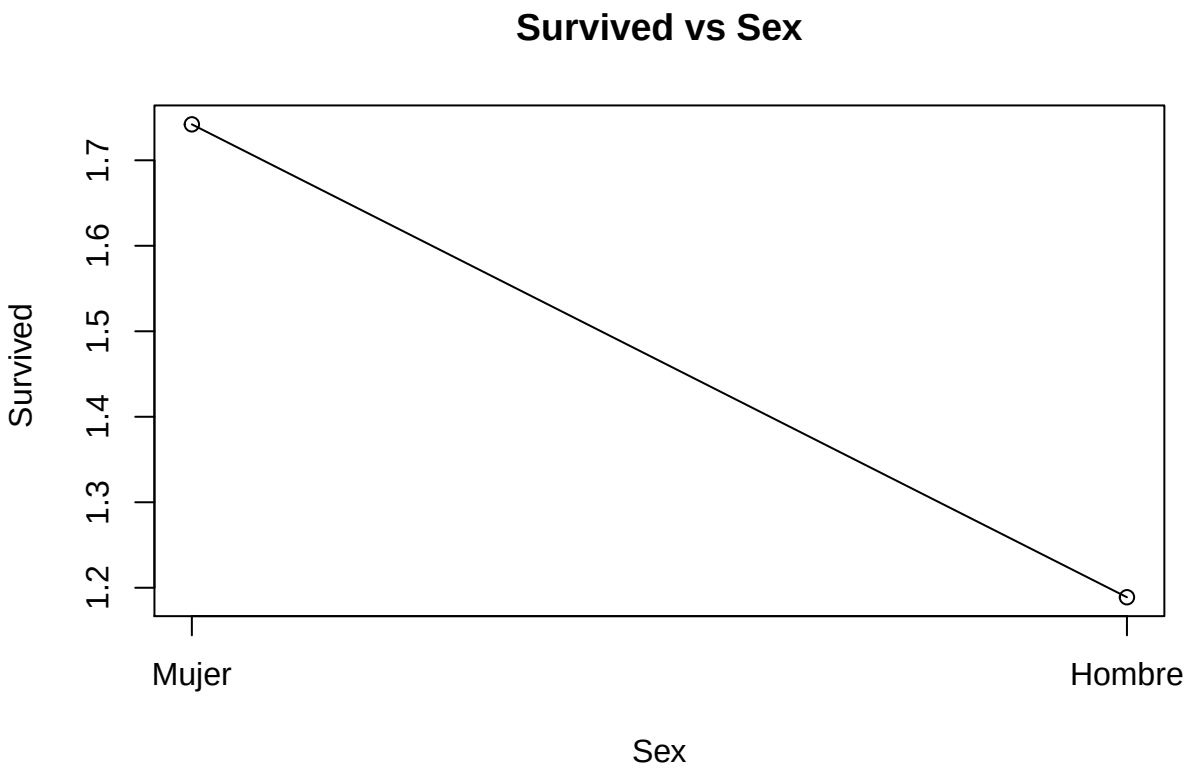
2.5. Representación de los resultados a partir de tablas y gráficas.

Para realizar una mejor representación grafica convertimos las variables categórico en numérico

```
df$Pclass = as.integer(df$Pclass)
df$Age1 = as.integer(df$Age1)
df$Sex = as.integer(df$Sex)
df$Survived = as.integer(df$Survived)
str(df)
```

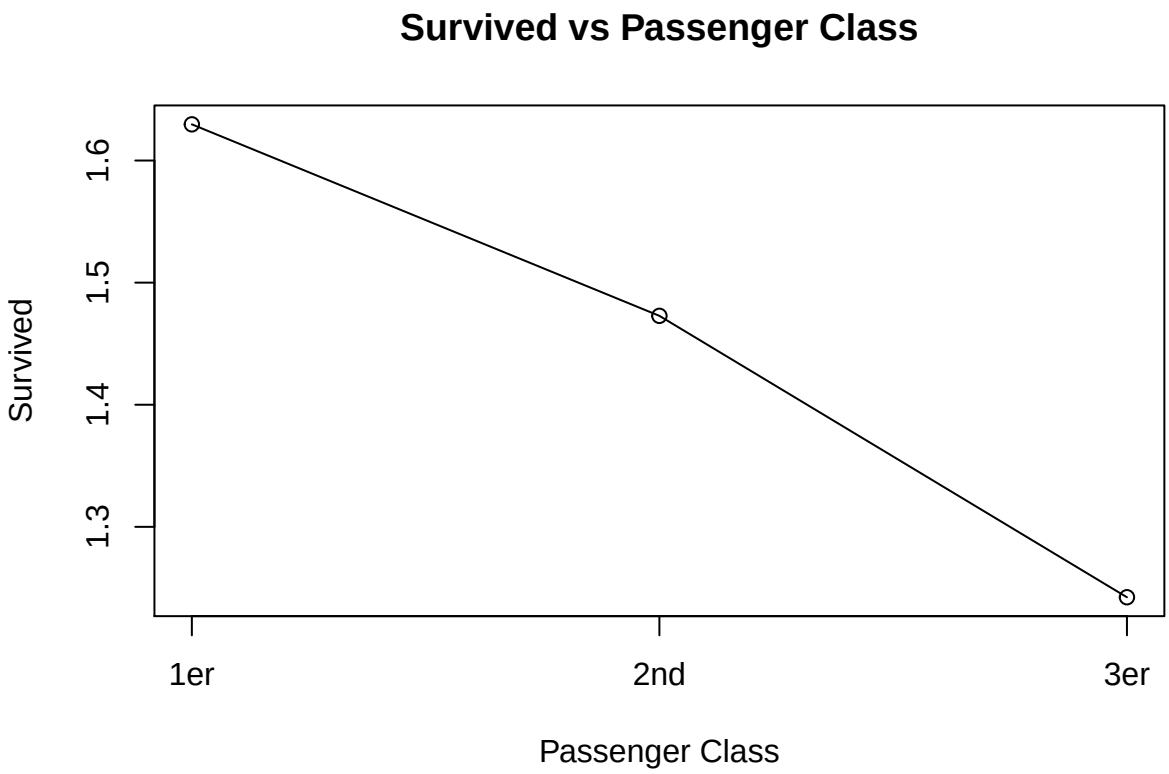
```
## 'data.frame':    891 obs. of  6 variables:
##  $ Survived: int  1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : int  2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Age1    : int  1 1 1 1 1 1 1 2 1 2 ...
```

```
mean_sex = c(0,0)
mean_sex[1] = mean(df$Survived[df$Sex==1])
mean_sex[2] = mean(df$Survived[df$Sex==2])
plot(mean_sex, type="o", main="Survived vs Sex", xlab="Sex", ylab="Survived ", xaxt="n")
axis(1, at=c(1,2), labels=c("Mujer", "Hombre"))
```



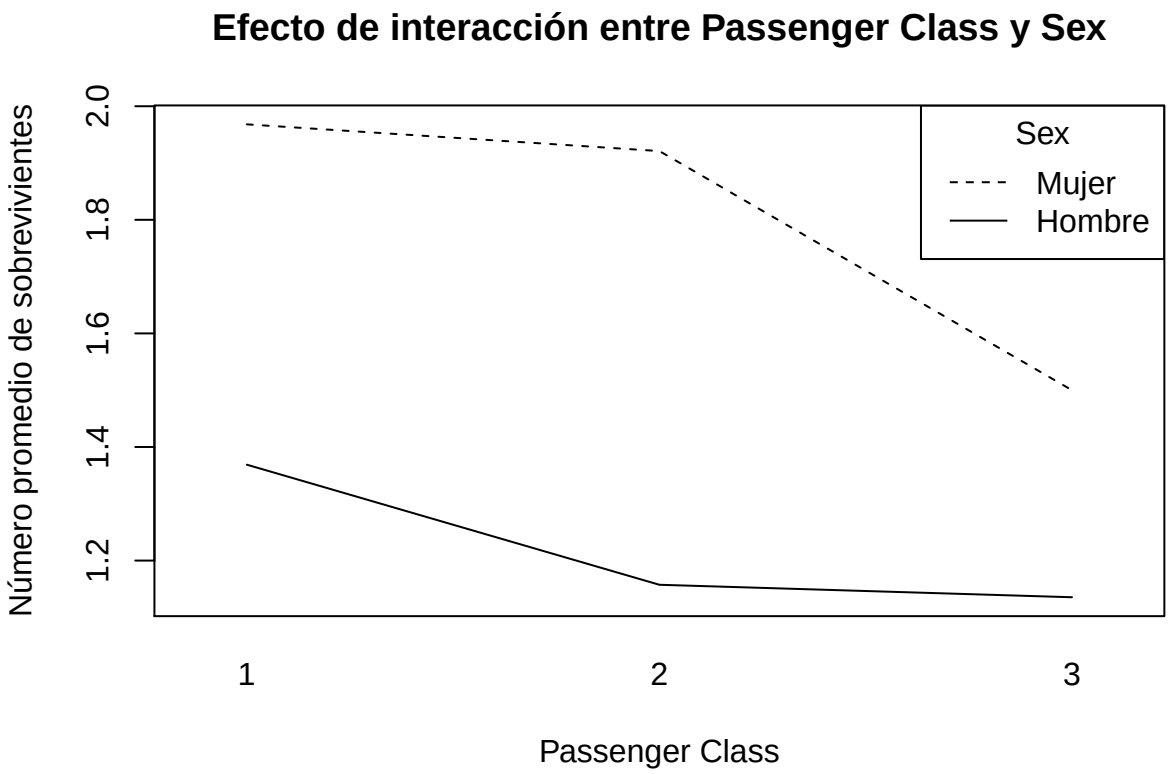

```
mean_pclass = c(0,0,0)
mean_pclass[1] = mean(df$Survived[df$Pclass==1])
mean_pclass[2] = mean(df$Survived[df$Pclass==2])
mean_pclass[3] = mean(df$Survived[df$Pclass==3])

plot(mean_pclass, type="o", main="Survived vs Passenger Class", xlab="Passenger Class", ylab="Survived"
axis(1, at=c(1,2,3), labels=c("1er", "2nd", "3er"))
```



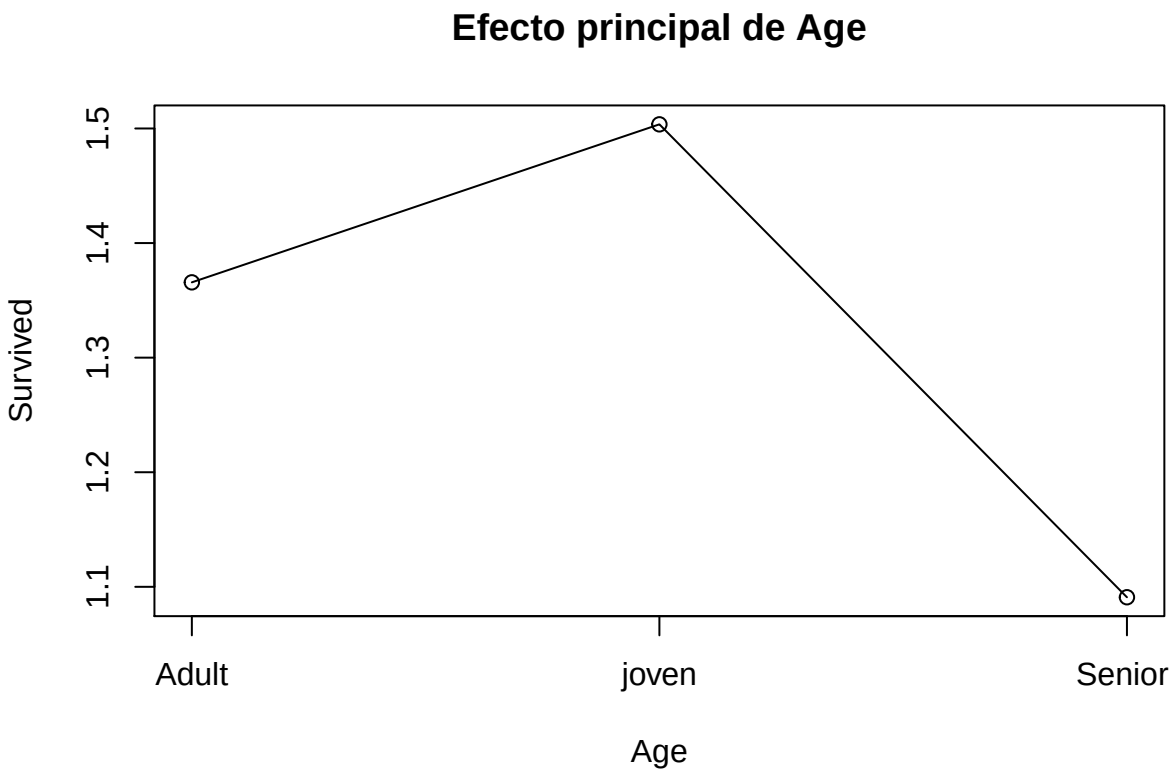
Observamos que el mayor número de sobrevivientes mujeres eran pasajeros de 1ra y 2da clase, en el caso de sobrevivientes hombres la mayoria de sobrevivientes pertenecen a 1ra clase.

```
interaction.plot(df$Pclass, df$Sex, df$Survived, xlab="Passenger Class", ylab="Número promedio de sobrev
                 main="Efecto de interacción entre Passenger Class y Sex", legend=FALSE)
legend("topright", c("Mujer","Hombre"), lty=c("dashed", "solid"), title="Sex")
```



Observamos que los supervivientes máximos en promedio de la categoría de edad eran jovenes (edad <=18 años), seguidos de adultos y personas mayores (edad > 65 años).

```
me_age = c(0,0,0)
me_age[1] = mean(df$Survived[df$Age1==1])
me_age[2] = mean(df$Survived[df$Age1==2])
me_age[3] = mean(df$Survived[df$Age1==3])
plot(me_age, type="o", main="Efecto principal de Age", xlab="Age", ylab="Survived", xaxt="n")
axis(1, at=c(1,2,3), labels=c("Adult", "joven", "Senior "))
```

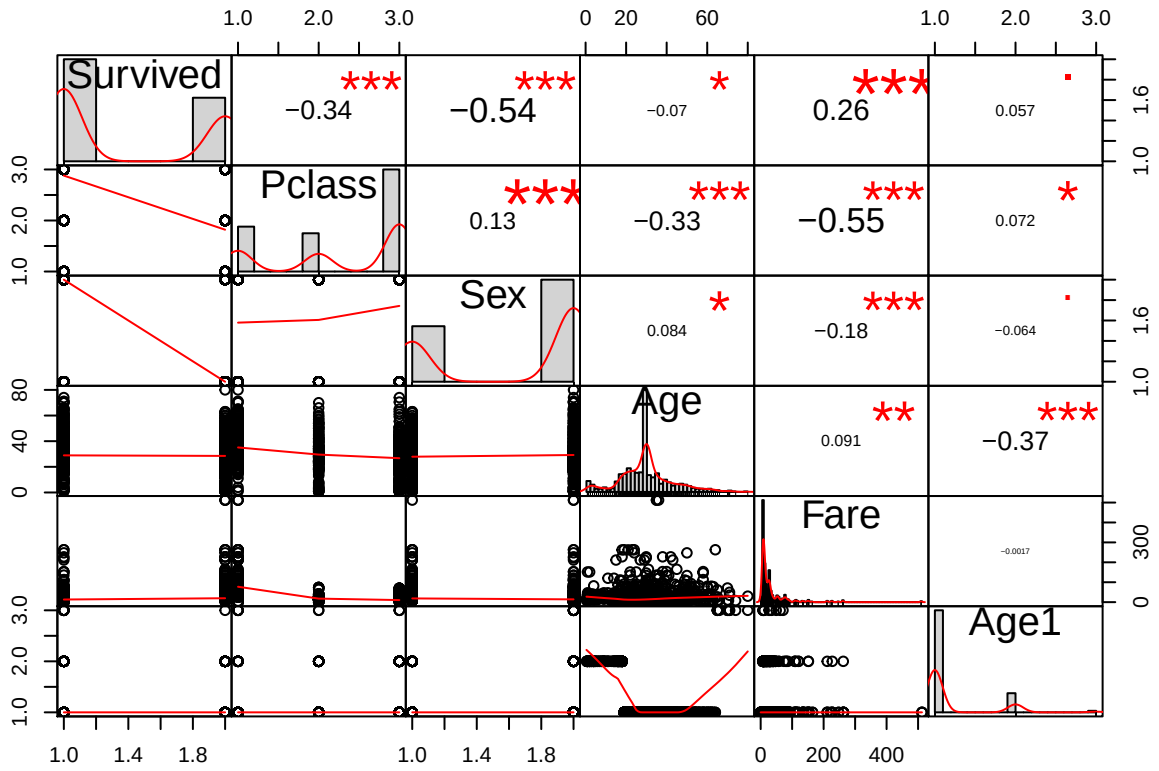


2.5.3.

```
chart.Correlation(df)
```

Cuadro 1: Correlación del conjunto de datos

	Survived	Pclass	Sex	Age	Fare	Age1
Survived	1.000	-0.338	-0.543	-0.070	0.257	0.057
Pclass	-0.338	1.000	0.132	-0.330	-0.549	0.072
Sex	-0.543	0.132	1.000	0.084	-0.182	-0.064
Age	-0.070	-0.330	0.084	1.000	0.091	-0.366
Fare	0.257	-0.549	-0.182	0.091	1.000	-0.002
Age1	0.057	0.072	-0.064	-0.366	-0.002	1.000




```
kable(cor(df), caption = "Correlación del conjunto de datos",digits = 3,  align = 'r')
```

Los valores mayores de 0,75 o menores de -0,75 son indicativos de una correlación alta positiva ó alta negativa. A partir de los resultados anteriores, las variables no están altamente correlacionadas en este conjunto de datos.

Como ejemplo probamos nuestro modelo de regresión logística para predecir la sobrevivencia de una mujer joven que viajó en 1ra clase pagando 200.

```
newd=data.frame(Age1= "joven" , Sex="female" , Pclass= "1" , Fare = 200 )

predict<- predict(modelo, newd, type= "response")
predict
```

```
##         1
```

```
## 0.9565746
```

La predicción de la probabilidad supervivencia para este caso es 0.9566.

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

A partir de los ánalisis realizados podemos indicar que los resultados permiten responder a los planteamientos iniciales del problema, conocer las variables con mayor influencia para la predicción de la probabilidad de supervivencia de los pasajeros del Titanic.

Se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos, como se ha visto las variables no fueron seleccionanas a priori ya que a lo largo del estudio, limpieza, ánalisis se ha ido evaluando el comportamiento de las variables. Para cada una de ellas, hemos podido ver que conocimientos aportan.

El análisis de correlación y el contraste de hipótesis nos han permitido conocer cuáles son las varibles que tienen influencia significativa sobre la probabilidad de suprevivencia de los pasajeros.

2.7. Código en R

El código de resolución de la práctica y el pdf de respuestas se encuentran en el repositorio GitHub, pueden ser accedidos a través de este enlace

2.7.1. Tabla de contibuciones al trabajo

Contribuciones	Firma
Investigación previa	J.A.L, N.N.S.P
Redacción de las respuestas	J.A.L, N.N.S.P
Desarrollo código	J.A.L, N.N.S.P