Tipología y ciclo de vida de los datos

Práctica 2 – Limpieza y análisis de datos

Francisco de Borja Navas Torres Álvaro de la Fuente Díaz

Junio 2021

Contents

1 Introducción **1**

2 Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder? **2**

3 Integración y selección de los datos de interés a analizar **2**

4 Limpieza de los datos **3**

4.1 Elementos vacíos 5

4.2 Valores extremos 6

5 Análisis de los datos **8**

5.1 Selección de los grupos de datos a analizar 8

5.2 Comprobación de la normalidad y homogeneidad de la varianza 9

5.3 Análisis estadístico del dataset 11

6 Representación gráfica de los resultados **15**

7 Conclusiones **17**

8 Contribuciones al trabajo **17**

---

1 Introducción

En esta segunda práctica de la asignatura de Tipología y ciclo de vida de los datos, vamos a tratar de limpiar y analizar los datos contenidos en el dataset 'titanic.csv'. Este dataset contiene información sobre sobre los pasajeros que abordaron en el Titanic desde los diferentes puertos en los que atracó.

---

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

El dataset elegido contiene información sobre el naufragio del Titanic. Estos datos nos permiten conocer, según una serie de características de los pasajeros que veremos a continuación, si el pasajero sobrevivió o no al accidente. El dataset costa de 891 ocurrencias para un total de 12 variables y lo hemos obtenido de la web de kaggle (https://www.kaggle.com/c/titanic).

Las variables contenidas en el dataset son las siguientes:

- PassengerId: identificador numérico del pasajero que embarcó en el Titanic
- Survived: Indicador de si el pasajero sobrevivió al accidente del Titanic.
  - 0: el pasajero no sobrevivió al accidente.
  - 1: el pasajero si sobrevivió al accidente.
- Pclass: Indicador del tipo de billete del pasajero.
  - 1: Primera clase.
  - 2: Segunda clase.
  - 3: Tercera clase.
- Name: nombre del pasajero, con el formato 'Apellido, Nombre'.
- Sex: sexo del pasajero. Los posibles valores son 'male' y 'female'.
- Age: edad del pasajero.
- SibSp: número de hermanos/cónyuge del pasajero a bordo.
- Parch: número de padres/hijos del pasajero a bordo.
- Ticket: identificador del billete.
- Fare: precio del billete
- Cabin: identificador del camarote del pasajero.
- Embarked: puerto en el que embarcó el pasajero:
  - C: Cherbourg
  - Q: Queenstown
  - S: Southampton

Con toda la información contenida en el conjunto de la muestra, tenemos como objetivo dar respuesta a qué variables del dataset tienen mayor influencia sobre la supervivencia del pasajero durante el conocido accidente del crucero Titanic. Con esta solución, podremos crear modelos predictivos que nos permitan conocer, dependiendo de las características de un hipotético pasajero, cuál es la probabilidad de que hubiese sobrevivido al accidente. Esta información no sólo es interesante desde un punto de vista 'curioso' sobre el accidente que ocurrió años atrás, sino que desde el punto de vista proyectivo puede ayudar al conocimiento de ingenieros navales para mejorar en el diseño y construcción de grandes transatlánticos que sean más seguros en accidentes de índole similar como alcances de grandes objetos sobre buques.

---

3 Integración y selección de los datos de interés a analizar

En primer lugar, cabe mencionar que el datset se encuentra dividido en dos ficheros, existe un fichero de entrenamiento y otro de testeo. El dataset está preparado para la ejecución de modelos. Por tanto, ya que nosotros no vamos a generar ningún modelo capaz de predecir la supervivencia de los pasajeros, solo vamos a utilizar el fichero de entrenamiento ya que tiene informado el atributo referente a la supervivencia.

Desde el punto de vista de la selección de atributos, descartamos aplicar cualquier técnica de filtrado ya que nos interesa realizar el análisis sobre el conjunto entero de ocurrencias. No vamos a discriminar ningún tipo de muestra dentro del dataset.

Por otro lado, no disponemos de ninguna variable que proceda de algún cálculo sobre otras, ni vemos relevante poder realizar cálculo alguno sobre más de una variable del dataset para intentar reducir el número de atributos sin que afecte en la representatividad de todas las variables en conjunto.

Basado en el análisis del total de atributos, si que vemos relevante descartar las variables PassengerId, Name y Ticket, puesto que son identificadores o descriptivos propios de cada uno de los pasajeros pero que no nos van a ayudar al análisis que nos compete, puesto que el valor de cada uno de estos atributos no presenta ningún tipo de agrupación sobre todos los demás.

---

4 Limpieza de los datos

El primer paso para poder realizar cualquier análisis sobre el dataset 'Titanic', es hacer la carga de los datos. Para ello, vamos a cargar el conjunto de datos almacenado en el fichero 'titanic_all.csv' y comprobar que R interpreta correctamente los tipos de los datos.

```
# Cargamos el fichero de datos
titanic_data <- read.csv('../data/titanic_train.csv',stringsAsFactors = FALSE)

# Primeros registros del dataset
head(titanic_data)
```

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

```
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
# Verificamos la estructura del conjunto de datos
str(titanic_data)
```

```
## 'data.frame':    891 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       : 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" ...
```

```
#Estadísticas básicas
summary(titanic_data)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

Podemos ver que el dataset está compuesto por 891 registros y 12 variables. En el apartado anterior, vimos que los atributos 'PassengerId', 'Name' y 'Ticket' no son relevantes de cara a nuestro estudio por lo que vamos a eliminarlos de la muestra.

```r
# Seleccionamos las variables relevantes
titanic_data <- titanic_data[c('Survived', 'Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Cabin', '
str(titanic_data)
```

```
## 'data.frame':    891 obs. of  9 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ 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 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin   : chr  "" "C85" "" "C123" ...
##  $ Embarked: chr  "S" "C" "S" "S" ...
```

4.1 Elementos vacíos

El primer paso del proceso de limpieza de los datos es el de detectar y corregir los posibles elementos vacíos. De esta forma nos aseguraremos que todos los registros del dataset están informados.

```r
# Valores perdidos
colSums(is.na(titanic_data))
```

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

```r
colSums(titanic_data=="")
```

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

Vemos que existen valores perdidos en los atributos 'Age', 'Embarked' y 'Cabin'. Para corregir la variable referente a la edad de los pasajeros, vamos a imputar los valores perdidos con la edad media de los valores presentes en la muestra. Los valores vacíos de las variables 'Embarked' y 'Cabin' los vamos a informar con el carácter 'U' referente a 'Unknown'.

```r
# Sustituimos los valores vacíos de la variable 'Age' por la media
titanic_data$Age[is.na(titanic_data$Age)] <- mean(titanic_data$Age,na.rm=T)

# Sustituimos los valores vacíos de las variable 'Embarked' y 'Cabin' por el valor U (unknown)
titanic_data$Embarked[titanic_data$Embarked==""]="U"
titanic_data$Cabin[titanic_data$Cabin==""]="U"
```

Comprobamos que se han corregido los valores perdidos.

```r
# Valores perdidos
colSums(is.na(titanic_data))
```

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

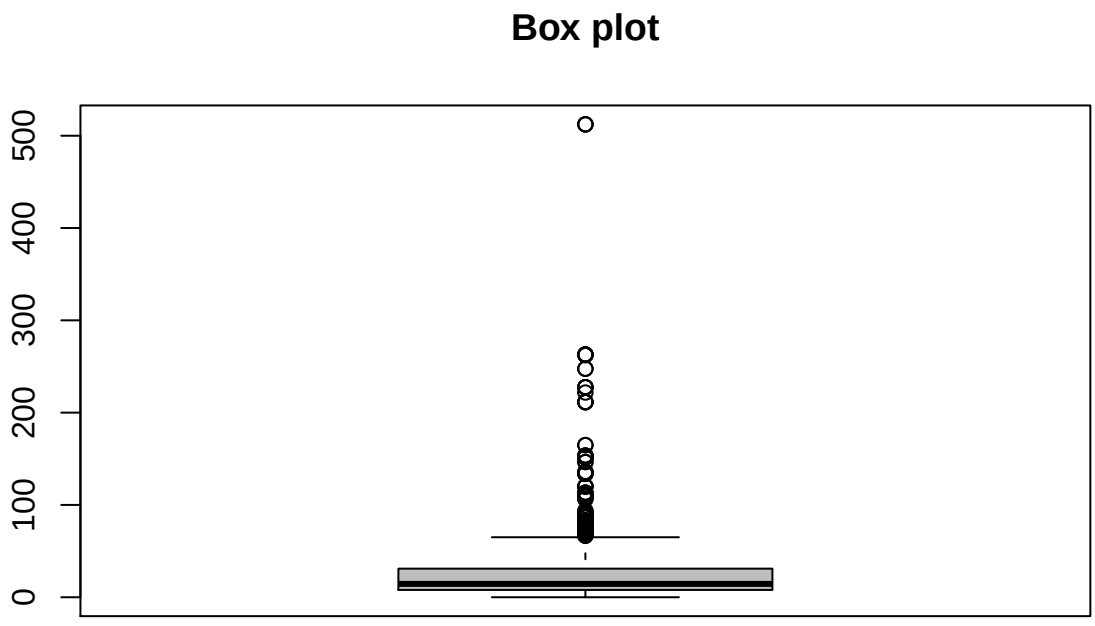
```r
colSums(titanic_data=="")
```

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

4.2 Valores extremos

Después de corregir los valores perdidos, vamos a analizar la posible existencia de valores extremos dentro del conjunto de la muestra. De entre los atributos del conjunto de los datos, solamente 'Age' y 'Fare' podrían tener valores extremos ya que son las únicas variables numéricas. Para ello, vamos a representar un diagrama de caja que nos permita identificar los posibles valores extremos.

```r
# Realizamos un diagrama de caja para comprobar si existen valores atípicos dentro de la variable 'Fare
boxplot(titanic_data$"Fare",main="Box plot", col="gray")
```

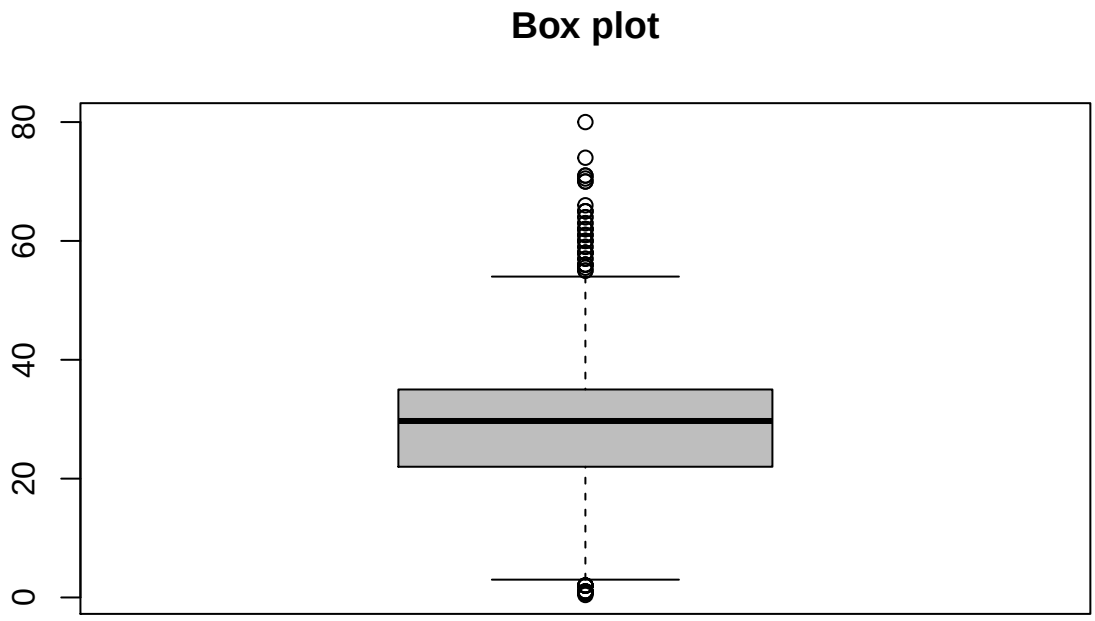


Vemos que existe al menos un valor que excede notablemente al resto con un valor superior a 500. En el otro extremo, existe algún pasajero que no pagó por embarcarse en el Titanic. Esto valor podría ser un valor centinela, asociado a casos excepcionales en los que por desconocimiento se inserta algún valor atípico. Estos valores extremos, pueden tener un impacto significativo a la hora de realizar el análisis de los datos por lo que vamos a imputarlos. Vamos a reemplazar estos valores por la media del resto de precios pagados por los pasajeros.

```
# Reemplazamos los valores extremo en la variable 'Fare' por la media del resto
mean_fare <- mean(titanic_data$Fare[(titanic_data$Fare < 500) & (titanic_data$Fare != 0)])
titanic_data$Fare[titanic_data$Fare > 500] <- mean_fare
titanic_data$Fare[titanic_data$Fare == 0] <- mean_fare
```

Una vez hemos corregido los valores extremos de la variable 'Fare', vamos a analizar la existencia de estos valores anómalos dentro del atributo 'Age'.

```
# Realizamos un diagrama de caja para comprobar si existen valores atípicos dentro de la variable 'Age'
boxplot(titanic_data$"Age",main="Box plot", col="gray")
```



Tal como podemos ver en el diagrama de caja anterior, no existen valores extremos. La edad mínima es un valor muy cercano a 0, lo que significa que en el transatlántico viajaba algún bebe mientras que la edad del pasajero más anciano era de 80 años.

Por último, una vez que hemos imputado los valores perdidos y hemos eliminado los valores extremos que pudieran falsear los resultados obtenidos, vamos a convertir las variables a categóricas para facilitar los análisis posteriores.

```r
# Convertimos la variable 'Sex' a categórica
titanic_data$Sex <- as.factor(titanic_data$Sex)

# Convertimos la variable 'Survived' a categórica
titanic_data$Survived <- as.factor(titanic_data$Survived)

# Convertimos la variable 'Pclass' a categórica
titanic_data$Pclass <- as.factor(titanic_data$Pclass)

# Convertimos la variable 'SibSp' a categórica
titanic_data$SibSp <- as.factor(titanic_data$SibSp)

# Convertimos la variable 'Parch' a categórica
titanic_data$Parch <- as.factor(titanic_data$Parch)

# Convertimos la variable 'Embarked' a categórica
titanic_data$Embarked <- as.factor(titanic_data$Embarked)
```

Extraemos unas estadísticas básicas del dataset para comprobar que todos los cambios se han realizado correctamente y que todas las variables tienen el tipo de dato correcto.

```r
#Estadísticas básicas
summary(titanic_data)
```

```
##  Survived Pclass      Sex           Age        SibSp   Parch
##  0:549    1:216   female:314   Min.   : 0.42   0:608   0:678
##  1:342    2:184   male  :577   1st Qu.:22.00   1:209   1:118
##           3:491                Median :29.70   2: 28   2: 80
##                                Mean   :29.70   3: 16   3:  5
##                                3rd Qu.:35.00   4: 18   4:  4
##                                Max.   :80.00   5:  5   5:  5
##                                                8:  7   6:  1
##       Fare            Cabin           Embarked
##  Min.   :  4.013   Length:891         C:168
##  1st Qu.:  7.925   Class :character   Q: 77
##  Median : 15.100   Mode  :character   S:644
##  Mean   : 31.108                      U:  2
##  3rd Qu.: 31.108
##  Max.   :263.000
##
```

---

5 Análisis de los datos

5.1 Selección de los grupos de datos a analizar

En este apartado vamos a seleccionar los grupos de datos presentes en el dataset, que vamos a utilizar durante el análisis posterior.

De entro de todos campos, vamos a seleccionar los atributos referentes al sexo del pasajero, la clase en la que viajaban, el puerto en el que embarcaron y si consiguieron sobrevivir al accidente.

```
# Agrupación por sexo
titanic_data.female <- titanic_data[titanic_data$Sex == 'female',]
titanic_data.male <- titanic_data[titanic_data$Sex == 'male',]

# Agrupación por clase
titanic_data.class_1 <- titanic_data[titanic_data$Pclase == '1',]
titanic_data.class_2 <- titanic_data[titanic_data$Pclass == '2',]
titanic_data.class_3 <- titanic_data[titanic_data$Pclass == '3',]

# Agrupación por supervivencia
titanic_data.survived_no <- titanic_data[titanic_data$Survived == '0',]
titanic_data.survived_yes <- titanic_data[titanic_data$Survived == '1',]

# Agrupación por el puerto de embarque
titanic_data.embarked_c <- titanic_data[titanic_data$Embarked == 'C',]
titanic_data.embarked_q <- titanic_data[titanic_data$Embarked == 'Q',]
titanic_data.embarked_s <- titanic_data[titanic_data$Embarked == 'S',]
titanic_data.embarked_u <- titanic_data[titanic_data$Embarked == 'U',]
```

5.2 Comprobación de la normalidad y homogeneidad de la varianza

Vamos a comprobar la normalidad de las variables cuantitativas presentes en la muestra. Para ello, vamos a emplear la prueba de normalidad de Anderson-Darling. Vamos a definir un valor alfa de 0,05 de manera que si el p-valor obtenido para cada variable en el análisis es superior al valor de alfa, podemos afirmar que esta variable sigue una distribución normal.

```
library(nortest)

alpha = 0.05
col.names = colnames(titanic_data)

for (i in 1:ncol(titanic_data)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(titanic_data[,i]) | is.numeric(titanic_data[,i])) {
    p_val = ad.test(titanic_data[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])
      cat(": p-valor ")
      cat(p_val)
      cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## Age: p-valor 3.7e-24
## Fare: p-valor 3.7e-24
```

Podemos ver como las variables 'Age' y 'Fare' tiene un p-valor de 3.7e-24. Este valor es muy inferior al valor alfa definido previamente por lo que podemos afirmar que no siguen una distribución normal.

Después de haber analizado la normalidad de las variables, vamos a estudiar la homogeneidad de la varianza utilizando el método Fligner-Killeen a través de las siguientes pruebas:

- Estudio de la homogeneidad del precio del billete en función del sexo del pasajero. Para ello, vamos a establecer la hipótesis nula de que ambas varianzas son iguales.

```
# Estudio de la homogeneidad del precio del billete en función del sexo del pasajero
fligner.test(Fare ~ Sex, data=titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Sex
## Fligner-Killeen:med chi-squared = 48.297, df = 1, p-value = 3.663e-12
```

El resultado de la prueba da un p-valor de 3.663e-12. Este valor es muy inferior a 0.05 por lo que rechazamos la hipótesis nula confirmando que ambas varianzas son diferentes.

- Estudio de la homogeneidad del precio del billete en función de si el pasajero sobrevivió al accidente. Para realizar el estudio, vamos a establecer la hipótesis nula de que ambas varianzas son iguales.

```
# Estudio de la homogeneidad del precio del billete en función de si el pasajero sobrevivió al accident
fligner.test(Fare ~ Survived, data=titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Survived
## Fligner-Killeen:med chi-squared = 80.298, df = 1, p-value < 2.2e-16
```

El p-valor obtenido al realizar la prueba es menor que 2.2e-16, por lo que al ser menor que 0.05 rechazamos la hipótesis nula. Es decir, las varianzas de ambos campos no son iguales.

- Estudio de la homogeneidad de la edad del pasajero en función de la clase en la que viajaba. Para ello, vamos a establecer la hipótesis nula de que ambas varianzas son iguales.

```
# Estudio de la homogeneidad de la edad del pasajero en función de la clase
fligner.test(Age ~ Pclass, data=titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Pclass
## Fligner-Killeen:med chi-squared = 34.97, df = 2, p-value = 2.55e-08
```

El resultado del análisis es un p-valor igual a 2.55e-08. Dado que este valor es inferior a 0.05, podemos afirmar que la hipótesis nula no es correcta confirmando que la varianza de las dos variables no es igual.

- Estudio de la homogeneidad de la edad del pasajero en función de su sexo. Para realizar este análisis vamos a establecer la hipótesis nula de que ambas varianzas son iguales.

```r
# Estudio de la homogeneidad de la edad del pasajero en función de su sexo
fligner.test(Age ~ Sex, data=titanic_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Sex
## Fligner-Killeen:med chi-squared = 1.052, df = 1, p-value = 0.305
```

En este caso, el resultado del test es un p-valor de 0.305. Al tratarse de un valor superior a 0.05 podemos afirmar que la hipótesis nula es correcta, confirmando que la varianza de ambas variables es igual.

5.3 Análisis estadístico del dataset

5.3.1 Modelo de regresión lineal

Como nuestra variable independiente supervivencia es dicotómica o binaria, debemos utilizar una regresión logística. A continuación, probaremos diferentes modelos, analizando la precisión de cada uno de ellos:

```r
# Nos guardamos en variables adicionales tanto las variables dependientes como las variables independie
Survived <- titanic_data$Survived
Pclass <- titanic_data$Pclass
Sex <- titanic_data$Sex
Age <- titanic_data$Age
SibSp <- titanic_data$SibSp
Parch <- titanic_data$Parch
Cabin <- titanic_data$Cabin
Embarked <- titanic_data$Embarked

# Creamos a continuación 5 modelos de regresión logarítmica
m1 <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Cabin + Embarked , data=titanic_data ,
m2 <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare , data=titanic_data , family="binomial")
m3 <- glm(Survived ~ Pclass + Sex + Age + Fare , data=titanic_data , family="binomial")
m4 <- glm(Survived ~ Pclass + Sex , data=titanic_data , family="binomial")
m5 <- glm(Survived ~ Age + Fare , data=titanic_data , family="binomial")

# Representamos las medidas Akaike Information Criterion para evaluar la precisión de los modelos
t.aic <- matrix(c(1, summary(m1)$aic,
2, summary(m2)$aic,
3, summary(m3)$aic,
4, summary(m4)$aic,
5, summary(m5)$aic),
ncol = 2, byrow = TRUE)
colnames(t.aic) <- c("MRL", "AIC")
t.aic
```

```
##      MRL       AIC
## [1,]   1  941.3640
## [2,]   2  805.9990
## [3,]   3  815.1154
## [4,]   4  834.8884
## [5,]   5 1123.7445
```

Nos quedamos con el modelo de menor AIC. Sería el modelo 2, que tiene en cuenta todas las variables dependientes indicadas en apartados anteriores a excepción de la cabina y el puerto de embarque. Hemos decidido probar con diferentes modelos sin tener en cuenta estas dos variables ya que por lógica desde el punto de vista de la supervivencia no considerábamos muy relevante el puerto donde embarcó el pasajero y la cabina. Esta última variable porque está informada para muy pocos pasajeros.

A continuación, interpretamos los resultados de la regresión:

```
summary(m2)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
##     Fare, family = "binomial", data = titanic_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7718  -0.6048  -0.4319   0.5911   2.4598
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.807e+00  4.890e-01   7.786 6.90e-15 ***
## Pclass2     -1.272e+00  3.133e-01  -4.062 4.88e-05 ***
## Pclass3     -2.268e+00  3.185e-01  -7.121 1.07e-12 ***
## Sexmale     -2.734e+00  2.008e-01 -13.616  < 2e-16 ***
## Age         -3.703e-02  8.324e-03  -4.449 8.64e-06 ***
## SibSp1       1.048e-01  2.244e-01   0.467  0.64040
## SibSp2      -2.131e-01  5.419e-01  -0.393  0.69413
## SibSp3      -2.203e+00  7.024e-01  -3.137  0.00171 **
## SibSp4      -1.794e+00  7.739e-01  -2.317  0.02048 *
## SibSp5      -1.607e+01  9.541e+02  -0.017  0.98656
## SibSp8      -1.595e+01  7.529e+02  -0.021  0.98310
## Parch1       4.365e-01  2.866e-01   1.523  0.12774
## Parch2       1.611e-01  3.848e-01   0.419  0.67536
## Parch3       2.971e-01  1.049e+00   0.283  0.77702
## Parch4      -1.591e+01  1.033e+03  -0.015  0.98772
## Parch5      -1.229e+00  1.169e+00  -1.051  0.29317
## Parch6      -1.656e+01  2.400e+03  -0.007  0.99449
## Fare        -1.185e-03  3.302e-03  -0.359  0.71979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance:  770.0  on 873  degrees of freedom
## AIC: 806
##
## Number of Fisher Scoring iterations: 15
```

Lo primero que observamos es que las variables que tienen un nivel de significancia inferior al 0.05 son 'Pclass', 'Sex', 'Age' y algunos valores de 'SibSp1'. A continuación, observando el signo de los coeficientes vemos que tanto para la clase, el sexo y la edad tiene signo negativo. Esto quiere decir que, a valores constantes del resto de variables, si aumentamos el valor de estas (las categóricas en su escala), la probabilidad de supervivencia

disminuye. A mayor edad menor supervivencia, si se trata de un hombre menor supervivencia y por último, a menor clase menor supervivencia.

A continuación, vamos a realizar una simulación de predicción de la supervivencia a través de unos datos inventados. De esta forma, vamos a testear el poder predictivo de nuestro modelo seleccionado.

```
# Datos inventados para realizar testeo
data <- data.frame(
Pclass = "1",
Sex = "female",
Age = 15,
SibSp = "1",
Parch = "1",
Fare= 500
)

predict(m2, data, type="response")
```

```
##         1
## 0.9608707
```

Observamos que para los datos introducidos tendríamos una probabilidad de sobrevivir al accidente del 96%.

5.3.2 Contraste de hipótesis

A continuación evaluaremos si la edad o el precio del billete influyen sobre la supervivencia al accidente.

5.3.2.1 Caso 1: Edad sobre la superviviencia. ¿Tienen mayor probabilidad de superviviencia las personas de menor edad?

Vamos a realizar un primer caso de estudio en el que vamos a tratar de relacionar la influencia que tiene la edad con la supervivencia del pasajero.

```
t.test(titanic_data.survived_yes$Age,titanic_data.survived_no$Age,
alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  titanic_data.survived_yes$Age and titanic_data.survived_no$Age
## t = -2.0385, df = 669.03, p-value = 0.02095
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.3581303
## sample estimates:
## mean of x mean of y
##  28.54978  30.41510
```

Obtenemos un p-valor inferior a nuestro coeficiente de significación del 0.05. Por tanto, rechazamos la hipótesis nula y afirmamos que la probabilidad de supervivencia será mayor cuanto menor sea la edad.

5.3.2.2 Caso 2: Precio del billete sobre la supervivivencia.¿Tienen mayor probabilidad de superviviencia las personas que gozan de un billete más exclusivo?

Después de comprobar como la edad sí que influyó en la supervivencia de los pasajeros, vamos a analizar si las personas con mayor nivel adquisitivo tuvieron mayor o menor tasa de supervivencia después del accidente.

```
t.test(titanic_data.survived_no$Fare,titanic_data.survived_yes$Fare,
alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  titanic_data.survived_no$Fare and titanic_data.survived_yes$Fare
## t = -7.0609, df = 506.65, p-value = 2.741e-12
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -16.37039
## sample estimates:
## mean of x mean of y
##  22.91116  44.26512
```

Obtenemos un p-valor inferior a nuestro coeficiente de significación del 0.05. Por tanto, rechazamos la hipótesis nula y afirmamos que la probabilidad de supervivencia será mayor cuanto mayor coste tenga el billete.

5.3.3 Análisis de correlaciones

A continuación, vamos a analizar la correlación existente entre las dos variables cuantitativas (Age y Fare). Para ello, vamos a utilizar el coeficiente de Spearman ya que los datos no siguen una distribución normal.

```
cor.test(titanic_data$Age, titanic_data$Fare, method="spearman")
```

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

```
##
##  Spearman's rank correlation rho
##
## data:  titanic_data$Age and titanic_data$Fare
## S = 102260916, p-value = 7.202e-05
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1325821
```

Vemos que el p-valor es significativo, y que el coeficiente de correlación está próximo a 0, por lo que podemos afirmar que estas variables tienen una correlación muy débil.

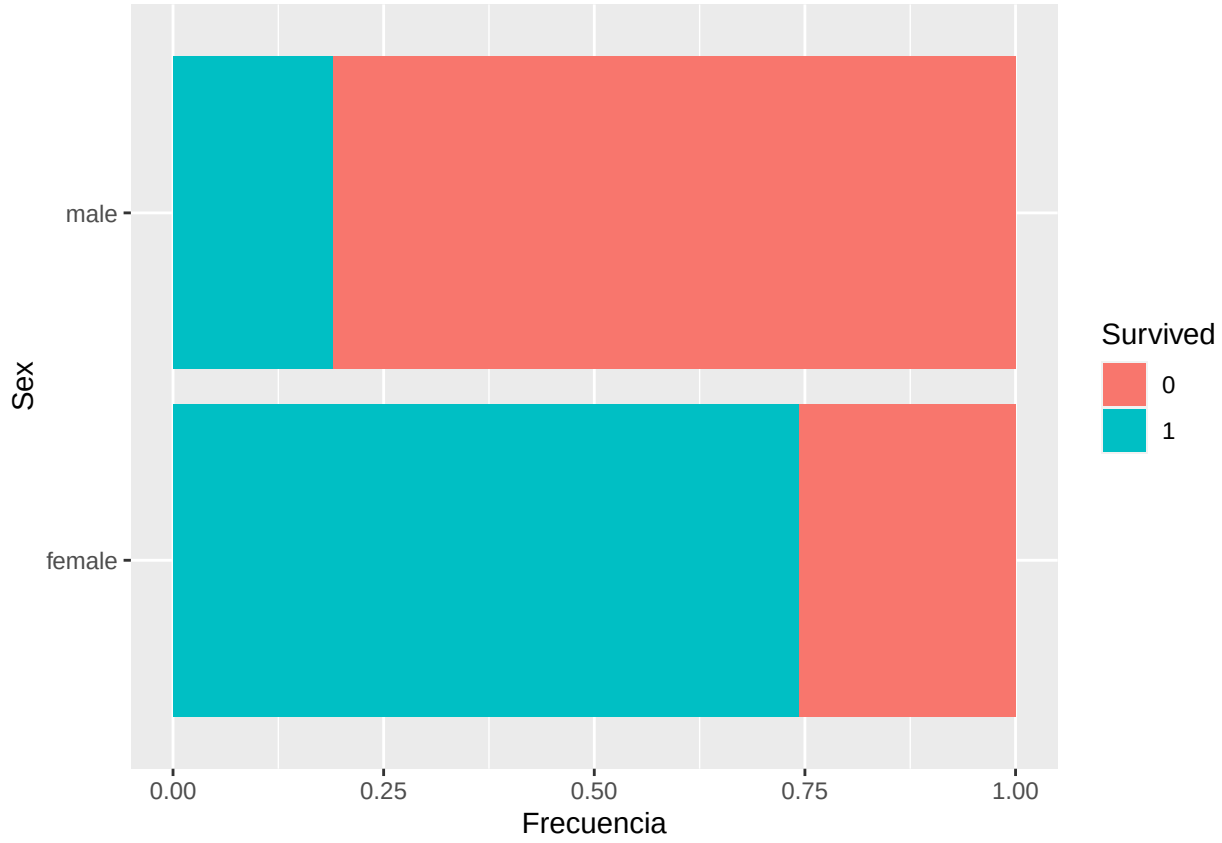
---

6 Representación gráfica de los resultados

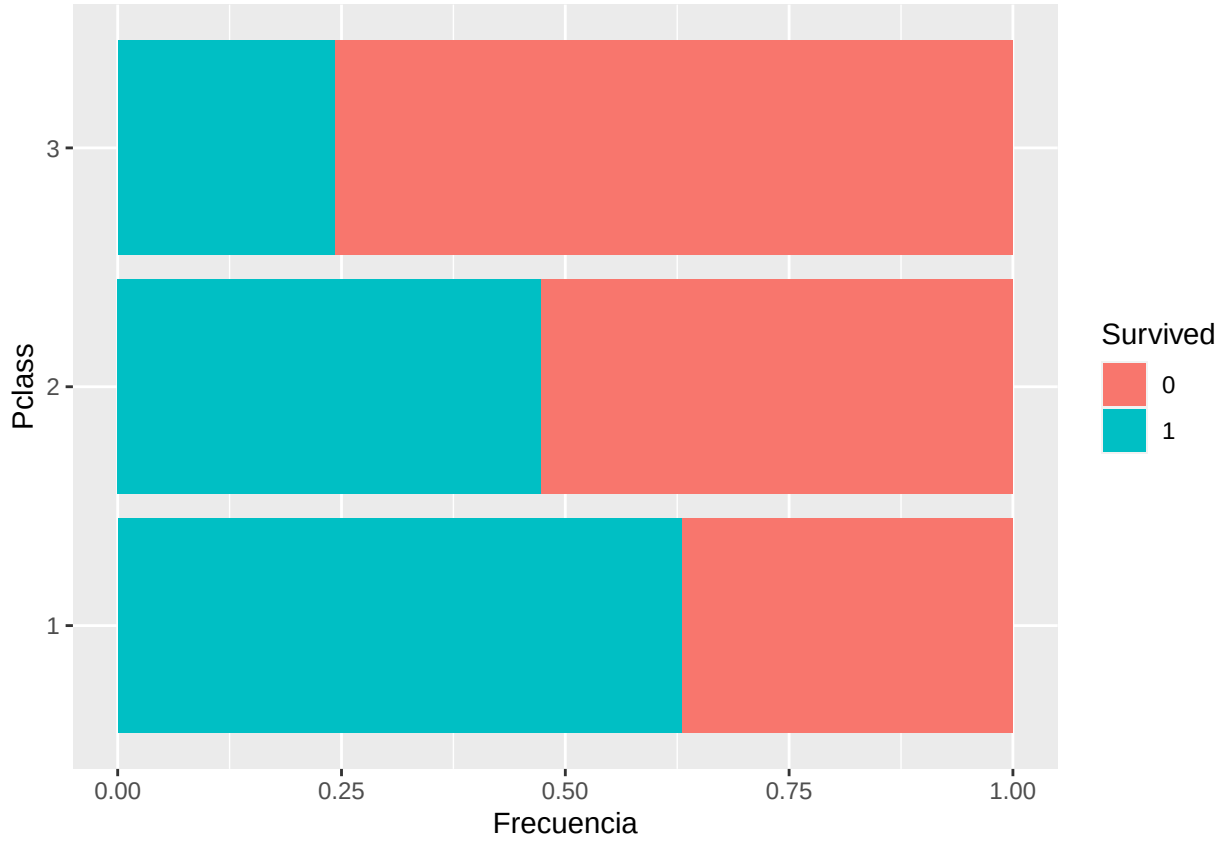
A continuación y con el objetivo de reforzar la información que hemos analizado en los pasos anteriores, haremos uso de visualizaciones mediante gráficos de barras e histogramas sobre la variable dependiente y las variables independientes más influyentes.

```
library(ggplot2)

# Diagrama de barras sexo-superviencia
ggplot(data = titanic_data,aes(y=Sex,fill=Survived))+geom_bar(position="fill")+xlab("Frecuencia")
```

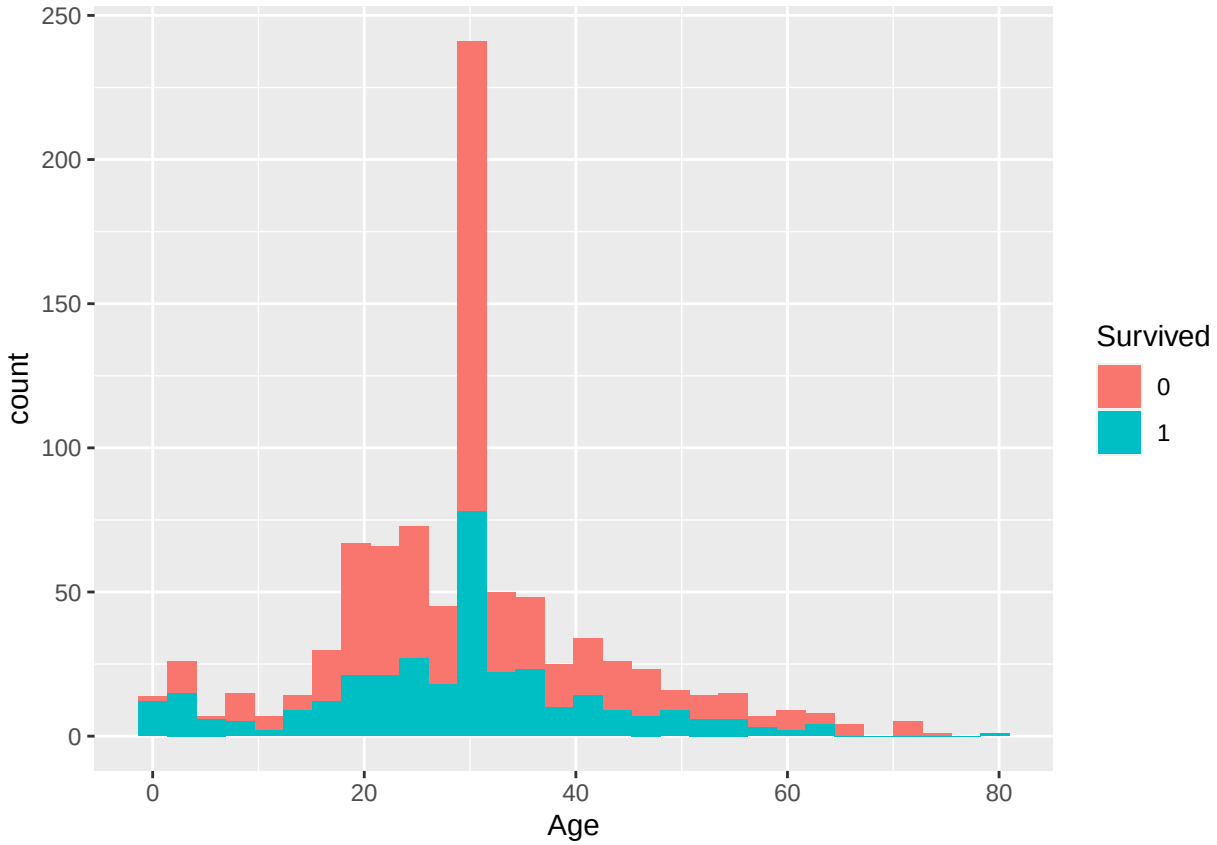



```
# Diagrama de barras clase-superviviencia
ggplot(data = titanic_data,aes(y=Pclass,fill=Survived))+geom_bar(position="fill")+xlab("Frecuencia")
```




```
# Histograma edad-supervivencia
ggplot(data = titanic_data,aes(x=Age,fill=Survived))+geom_histogram()
```

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



7 Conclusiones

En base a los resultados obtenidos durante todo el análisis de los datos contenidos en el dataset, hemos observado que:

- Las mujeres tuvieron mayor índice de supervivencia tras el accidente, por lo que ser de este sexo implica tener mayor probabilidad de sobrevivir.
- La clase en la que viajaban los pasajeros tiene una relación directa con la supervivencia de la persona. De esta forma, cuanto mayor sea la clase del billete, mayor probabilidad de supervivencia.
- Otro aspecto destacable es el relacionado con la edad, las personas jóvenes tuvieron mayor supervivencia tras el impacto con el iceberg. Es decir, cuanto más joven, mayor probabilidad de supervivencia.

8 Contribuciones al trabajo

En la siguiente tabla se detallan los nombres de las personas que han participado en el desarrollo de las tareas de análisis y limpieza de los datos, así como en la elaboración de las respuestas a las preguntas planteadas en el enunciado de la actividad y contenidas en el este fichero pdf.

Contribuciones	Firma
Investigación previa	FBNT y AFD
Redacción de las respuestas	FBNT y AFD
Desarrollo código	FBNT y AFD