PRA2 - Proyecto Analítico

2022-05-16

Se instalan y cargan las librerías.

```
# https://cran.r-project.org/web/packages/ggplot2/index.html
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
# https://cran.r-project.org/web/packages/dplyr/index.html
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('caret')) install.packages('caret'); library('caret')
if (!require('normtest')) install.packages('normtest'); library('normtest')
if (!require('car')) install.packages('car'); library('car')
if (!require('gridExtra')) install.packages('gridExtra'); library('gridExtra')
```

Ejercicio 1

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

1.1 Carga de los datos

Se carga el dataset.

```
df <- read.csv('../data/train.csv',stringsAsFactors = FALSE)
```

Se confirma que el dataset se ha cargado correctamente:

```
head(df)
```

```
##   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
```

1.2 Descripción del dataset

Descripción del dataset:

```
str(df)
```

```
## '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" ...
```

Se observa que el conjunto de datos consta de 891 observaciones y 12 variables que describen la tripulación del Titanic.

PassengerId Número identificador del pasajero (Integer).

Survived Variable que especifica si un pasajero ha sobrevivido. 0 = No, 1 = Si (Integer).

Pclass Clase socio económica del pasajero. 1 = Clase Alta, 2 = Clase Media, 3 = Clase baja (Integer).

Name Nombre del pasajero (String).

Sex Sexo del pasajero (String).

Age Edad del pasajero. Para pasajeros de menos de un año, la edad esta fraccionada. Si la edad esta estimada, su formato es xx.5 (Numeric).

SibSp Número de familiares, hermanos o esposos, a abordo (Integer).

Parch Número de familiares, padres o hijos, a bordo (Ingeger).

Ticket Número del billete (String).

Fare Precio del billete (Numeric).

Cabin Número de cabina (String).

Embarked Puerto en el que embarco el pasajero. C = Cherbourg, Q = Queenstown, S = Southampton (String).

Se observa que muchas de estas variables se presentan como Integers o Strings, cuando realmente nos expresan una categoría. Por ello, se factorizarán las variables categoricas.

```
df$Survived <- factor(df$Survived)
df$Pclass <- factor(df$Pclass)
df$Sex <- factor(df$Sex)
df$SibSp <- factor(df$SibSp)
df$Parch <- factor(df$Parch)
df$Embarked <- factor(df$Embarked)
```

1.3 Motivación del estudio

Cuando el Titanic fue construido, se considero que este era imposible de hundir. Durante su viaje inaugural, el Titanic se hundió tras chocar contra un iceberg y la escasez de barcos salvavidas convirtió el accidente en una gran tragedia que resultó en la muerte de 1502 personas del conjunto de 2224 pasajeros y tripulación. Claramente, la suerte tuvo un papel importante en determinar quien sobrevivió y quien no, pero se puede intuir que ciertos grupos de personas tuvieron más probabilidades de hacerlo que otros.

Este estudio se centra en responder a la pregunta "¿Qué grupos de personas tuvieron más probabilidades de sobrevivir el accidente?", o dicho de otra forma, "¿Cuáles fueron los factores de peso que influyeron en que los pasajeros tuvieran mayor probabilidad de sobrevivir?"

Ejercicio 2

Integración y selección de los datos de interés a analizar. Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

La principal variable a incluir va a ser **Survived**. Esta será la variable objetivo a estudiar ya que se quiere determinar que factores tuvieron más peso al aumentar la probabilidad de sobrevivir de los pasajeros. A continuación, se seleccionan todas las variables que puedan describir el pasajero, y se eliminan las variables irrelevantes. Si se observa la previa descripción de las variables realizada en el ejercicio 1, rápidamente se puede identificar que las variables **Cabin**, el número de cabina, y **Ticket**, el número del billete, no aportan información relevante para describir al pasajero. Del mismo modo, las variables **Nombre** y **PassangerId** que sirven para identificar a los pasajeros, tampoco aportan información descriptiva sobre ellos.

Se procede a descartar las variables irrelevantes.

```
irrelevant <- c("Cabin", "Ticket", "Name", "PassengerId")
titanic <- df[ , -which(names(df) %in% irrelevant)]
```

Ejercicio 3

Limpieza de los datos.

3.1 Valores vacios

¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.

Estadísticas de valores vacíos.

```
colSums(is.na(titanic))
```

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

```
colSums(titanic=="")
```

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

Se observan 177 valores vacíos en la variable **Age** y 2 valores vacíos en la variable **Embarked**.

Centrando la atención en el caso de la variable **Age**, esta contiene 177 valores vacíos. Como eliminar 177 observaciones del conjunto de datos supondría una gran perdida de información ya que este contiene 801 observaciones, se opta por la estrategia de completar los valores vacíos usando la media de edad del conjunto de datos.

```
titanic$Age[is.na(titanic$Age)] <- mean(titanic$Age, na.rm=T)
```

Debido a que la variable **Embarked** tiene solo 2 valores vacíos, no se ve la necesidad de crear una categoría nueva "Desconocido" para substituir los valores vacíos. La mejor estrategia en este caso es sustituir estos valores vacíos por la categoría más repetida, evitando así la perdida de información que supondría eliminar las observaciones.

```
# Determinamos la categoría que se repite con mayor frecuencia
table(titanic$Embarked)
```

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

Se observa que la categoría 'S' es la que se repite con mayor frecuencia.

Se procede a sustituir los valores vacíos por 'S'.

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

Se comprueba que ya no existen valores vacíos en el conjunto de datos.

```
colSums(is.na(titanic))
```

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

```
colSums(titanic=="")
```

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

3.2 Valores extremos

Identifica y gestiona los valores extremos.

Los valores extremos se dan en las variables continuas cuando se observan valores significativamente altos o significativamente bajos, que generan duda sobre su veracidad.

A continuación se muestran las estadísticas básicas del conjunto de datos para identificar posibles valores extremos (*outliers*) en las variables continuas **Age**, **SibSp**, **Parch** y **Fare**.

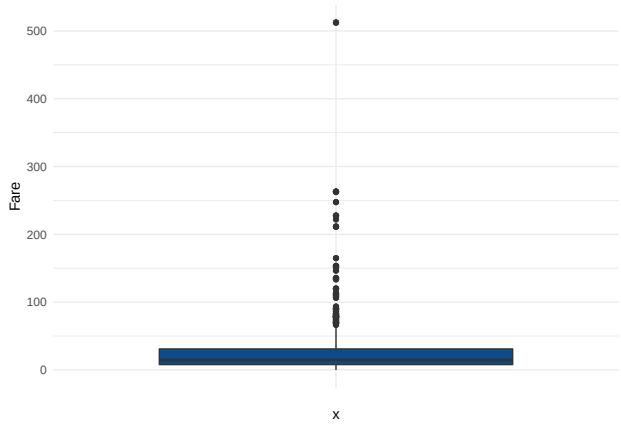
```
summary(titanic)
```

```
##  Survived Pclass      Sex           Age        SibSp   Parch        Fare
##  0:549    1:216   female:314   Min.   : 0.42   0:608   0:678   Min.   :  0.00
##  1:342    2:184   male  :577   1st Qu.:22.00   1:209   1:118   1st Qu.:  7.91
##           3:491                Median :29.70   2: 28   2: 80   Median : 14.45
##                                Mean   :29.70   3: 16   3:  5   Mean   : 32.20
##                                3rd Qu.:35.00   4: 18   4:  4   3rd Qu.: 31.00
##                                Max.   :80.00   5:  5   5:  5   Max.   :512.33
##                                                8:  7   6:  1
##  Embarked
##   :  0
##  C:168
##  Q: 77
##  S:646
##
##
##
```

Se observa que los valores mínimos y máximos de la variable **Age**, se mantienen dentro de un rango lógico. **SibSp** y **Parch**, tienen un valor máximo alto, pero parece consistente con los estándares de familias de la época. La variable que será necesaria analizar con profundidad es **Fare**. Esta variable contiene un valor máximo muy superior al valor medio.

Se representa el *boxplot* para identificar los *outliers* de **Fare**.

```
ggplot(titanic) +
  aes(x = "", y = Fare) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



Con esta representación, se pueden identificar claramente *outilers* correspondientes al valor máximo de 512.33. Se opta por eliminar estos *outliers*, ya que la diferencia respecto a las demás observaciones es tan significativa que se sospecha que tales observaciones contienen información no verídica. Los otros puntos que son detectados como *outliers*, menores a 300, corresponden a los precios pagados por los pasajeros de primera clase. A continuación, se muestran algunas de estas observaciones para **Fare** mayor a 100.

```
head(titanic[titanic$Fare > 150, ])
```

```
##     Survived Pclass    Sex Age SibSp Parch     Fare Embarked
## 28         0      1   male  19     3     2 263.0000        S
## 89         1      1 female  23     3     2 263.0000        S
## 119        0      1   male  24     0     1 247.5208        C
## 259        1      1 female  35     0     0 512.3292        C
## 269        1      1 female  58     0     1 153.4625        S
## 298        0      1 female   2     1     2 151.5500        S
```

Se observa que efectivamente, esas observaciones corresponden a pasajeros de primera clase.

Eliminamos los *outliers* detectados:

```
titanic <- subset(titanic, titanic$Fare < 500)
```

3.3 Almacenamiento del dataset final

Guardamos el conjunto de datos limpio.

```
write.csv(titanic, "../data/Titanic_clean.csv", row.names = FALSE)
```

Ejercicio 4

Análisis de los datos.

4.1 Elección de los datos a analizar

En este análisis se pretende analizar el peso de las distintas variables del conjunto sobre la variable dependiente **Survived**. En el Ejercicio 2, se escogió el subset del conjunto de datos considerado de interés analítico para el estudio. Este consiste de las variables **Survived**, **Pclass**, **Sex**, **Age**, **SibSp**, **Parch**, **Fare** y **Embarked**.

De las variables de interés para el análisis, se crea una agrupación por genero, la variable **Sex**, para realizar un análisis posteriormente por contraste de hipótesis.

```
# Agrupación por género
titanic.men   <- subset(titanic, Sex == "male")
titanic.women <- subset(titanic, Sex == "female")
```

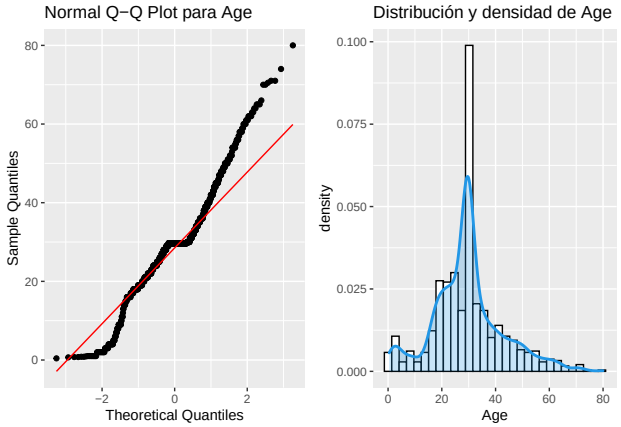
4.2 Normalidad y Homogeneidad

Comprobación de la normalidad y homogeneidad de la varianza.

Se analiza la normalidad y la homogeniedad de la varianza para las variables continuas **Age** y **Fare**.

Age: Análisis visual de normalidad:

```
p1 <- ggplot(titanic, aes(sample = Age)) + stat_qq() +
        stat_qq_line(colour= "red") +
        ylab("Sample Quantiles") + xlab("Theoretical Quantiles") +
        ggtitle("Normal Q-Q Plot para Age")

#[1]
p2 <- ggplot(titanic, aes(x = Age)) +
        geom_histogram(aes(y = ..density..), bins = 30,
                       colour = 1, fill = "white") +
        geom_density(lwd = 1, colour = 4,
                     fill = 4, alpha = 0.25) +
        ggtitle("Distribución y densidad de Age")

grid.arrange(p1,p2,ncol = 2, nrow = 1)
```



Para determinar que puede haber normalidad, se debería poder ver como la mayoría de observaciones se alinea encima de la línea roja (indica el comportamiento esperado de una variable normal) para la representación Quantil-Quantil. Este no es el caso observado para la variable **Age**.

Centrando la atención en el histograma de la variable con su correspondiente curva de densidad, para que hubiera normalidad, se debería observar una curva de densidad simétrica con forma de campana, por lo que se puede determinar que no es una variable normal. Se puede observar de la distribución, que la mayoría de pasajeros eran jóvenes de entre 20 y 30 años.

Análisis por tests de normalidad:

```
#[2]
## Shapiro test
shapiro.test(titanic$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Age
## W = 0.95851, p-value = 3.678e-15
```

```
## Jarque Bera test
jb.norm.test(titanic$Age)
```

```
##
##  Jarque-Bera test for normality
##
## data:  titanic$Age
## JB = 61.607, p-value < 2.2e-16
```

```
## Forsini test
frosini.norm.test(titanic$Age)
```

```
##
##  Frosini test for normality
##
## data:  titanic$Age
## B = 1.4197, p-value < 2.2e-16
```

Para los tests de normalidad Shapiro-Wilk, Jarque-Bera y Frosini, la hipótesis nula indica normalidad, mientras que la hipótesis alternativa indica que no hay normalidad. Un valor p superior al nivel de significancia ~0.05 indicaría que no se puede descartar la hipótesis nula de normalidad. Para todos los test probados, se observan unos valores p muy inferiores al nivel de significancia, que indica que no estamos frente a una variable normal.

Análisis de homogeniedad de la varianza:

```
leveneTest(y = titanic$Age, group = titanic$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value  Pr(>F)
## group   1  5.7214 0.01697 *
##       886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para evaluar la homocedasticidad, o la homogeneidad de la varianza, para la variable **Age**, se debe tener en cuenta que, como se determinó previamente, no se trata de una variable normal. Por ello, la elección del test Levene es la más conveniente siendo este menos sensible a que la ausencia de normalidad [3].

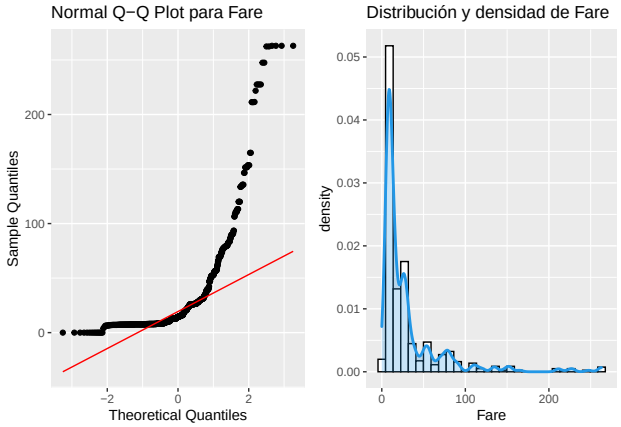
Hipótesis nula: Ambas varianzas son iguales. Hipótesis alternativa: Las varianzas son distintas entre si.

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

Se aplica el test para la variable **Age** y se comparan las varianzas cuando esta variable es agrupada según **Survived**, la variable dependiente que se quiere estudiar. Se puede observar de los resultados un valor p igual a 0.0169. Esto indica que se debe descartar la hipótesis nula, es decir, el hecho de que ambas varianzas sean iguales, en el caso de optar por un nivel de significancia de 0.01 o 0.001. Si se opta por un nivel de significancia de 0.05 o 0.1, no se puede descartar la hipótesis nula y por lo tanto las varianzas de ambos grupos deben considerarse iguales.

Fare Análisis visual de normalidad:

```
p1 <- ggplot(titanic, aes(sample = Fare)) + stat_qq() +
        stat_qq_line(colour= "red") +
        ylab("Sample Quantiles") + xlab("Theoretical Quantiles") +
        ggtitle("Normal Q-Q Plot para Fare")

#[1]
p2 <- ggplot(titanic, aes(x = Fare)) +
        geom_histogram(aes(y = ..density..), bins = 30,
                       colour = 1, fill = "white") +
        geom_density(lwd = 1, colour = 4,
                     fill = 4, alpha = 0.25) +
        ggtitle("Distribución y densidad de Fare")
grid.arrange(p1,p2,ncol = 2, nrow = 1)
```



Se observa en el gráfico Normal Q-Q, como la mayoría de observaciones no se alinea encima de la línea roja que indica como debería comportarse una variable normal. Este gráfico indica que **Fare** no sigue una distribución normal.

En el histograma de la variable con su curva de densidad, se ve como la mayoría de valores se centran entre 0 y 50, con un pico pronunciado entorno a 10, y decaen casi exponencialmente. La curva de densidad no es simétrica ni tiene forma de campana. Se determina que no se trata de una variable con distribución normal.

Análisis por tests de normalidad:

```
#[2]
## Shapiro test
shapiro.test(titanic$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic$Fare
## W = 0.60472, p-value < 2.2e-16
```

```
## Jarque Bera test
jb.norm.test(titanic$Fare)
```

```
##
##  Jarque-Bera test for normality
```

```
##
## data:  titanic$Fare
## JB = 6793.2, p-value < 2.2e-16
```

```
## Forsini test
frosini.norm.test(titanic$Fare)
```

```
##
##  Frosini test for normality
##
## data:  titanic$Fare
## B = 3.9484, p-value < 2.2e-16
```

Para los tests de normalidad Shapiro-Wilk, Jarque-Bera y Frosini, la hipótesis nula indica normalidad, mientras que la hipotesis alternativa indica que no hay normalidad. En todos aparece un valor p muy pequeño, inferior a todos los limites de significancia comunmente usados, por lo que se debe descartar la hipótesis nula de normalidad.

Análisis de homogeniedad de la varianza:

```
leveneTest(y = titanic$Fare, group = titanic$Survived, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##        Df F value    Pr(>F)
## group   1  44.703 4.055e-11 ***
##       886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para evaluar la homocedasticidad, o la homogeniedad de la varianza, para la variable **Fare**, se elige de nuevo el test Levene siendo este menos sensible a que la ausencia de normalidad [3].

Hipótesis nula: Ambas varianzas son iguales. Hipótesis alternativa: Las varianzas son distintas entre si.

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

Se observa que al evaluar las varianzas agrupando la variable **Fare** en función de la variable dependiente **Survived**, estas son claramente diferentes entre ellas. Se obtiene un valor p muy pequeño, inferior a todos los límites de significancia comúnmente usados, y por ello se descarta las hipótesis nula.

4.3 Análisis

Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

4.3.1 Analisis de la relación entre haber sobrevivido, o no, con las variables categóricas

4.3.1.1 Prueba Chi-cuadrado Para confirmar si hay relación entre las variables categóricas **Sex**, **Pclass**, **SibSp**, **Parch** y **Embarked** con la variable **Survive** que indica si los pasajeros sobrevivieron o no al accidente, se utilizará el test Chi-cuadrado de Pearson [4].

Para el test Chi-cuadrado, las hipótesis son las siguientes:

Hipótesis nula: No existe relación entre ambas variables. Hipótesis alternativa: Existe relación entre las variables.

Un valor p inferior a los valores comunes de significancia ~0.05, indica que se puede descartar la hipótesis nula, por lo tanto, existe relación, de lo contrario, indicaría que no hay relación entre ambas variables.

Sex:

```
with(titanic, chisq.test(Survived, Sex))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  Survived and Sex
## X-squared = 262.28, df = 1, p-value < 2.2e-16
```

Hay relación.

Pclass:

```
with(titanic, chisq.test(Survived, Pclass))
```

```
##
##  Pearson's Chi-squared test
##
## data:  Survived and Pclass
## X-squared = 100.03, df = 2, p-value < 2.2e-16
```

Hay relación.

SibSp:

```
with(titanic, chisq.test(Survived, SibSp))
```

```
## Warning in chisq.test(Survived, SibSp): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Survived and SibSp
## X-squared = 37.981, df = 6, p-value = 1.133e-06
```

Hay relación, pero se observa un aviso que indica que existe cierto error en el resultado. Este aviso es provocado por categorias con muy pocas observaciones[5], lo evaluaremos posteriormente con un análisis visual.

Parch:

```
with(titanic, chisq.test(Survived, Parch))
```

```
## Warning in chisq.test(Survived, Parch): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Survived and Parch
## X-squared = 27.665, df = 6, p-value = 0.0001087
```

Hay relación, pero se obsrva un aviso que indica que existe cierto error en el resultado. Este aviso es provocado por categorias con muy pocas observaciones[5], evaluará posteriormente con un análisis visual.

Embarked:

```
with(titanic, chisq.test(Survived, Embarked))
```

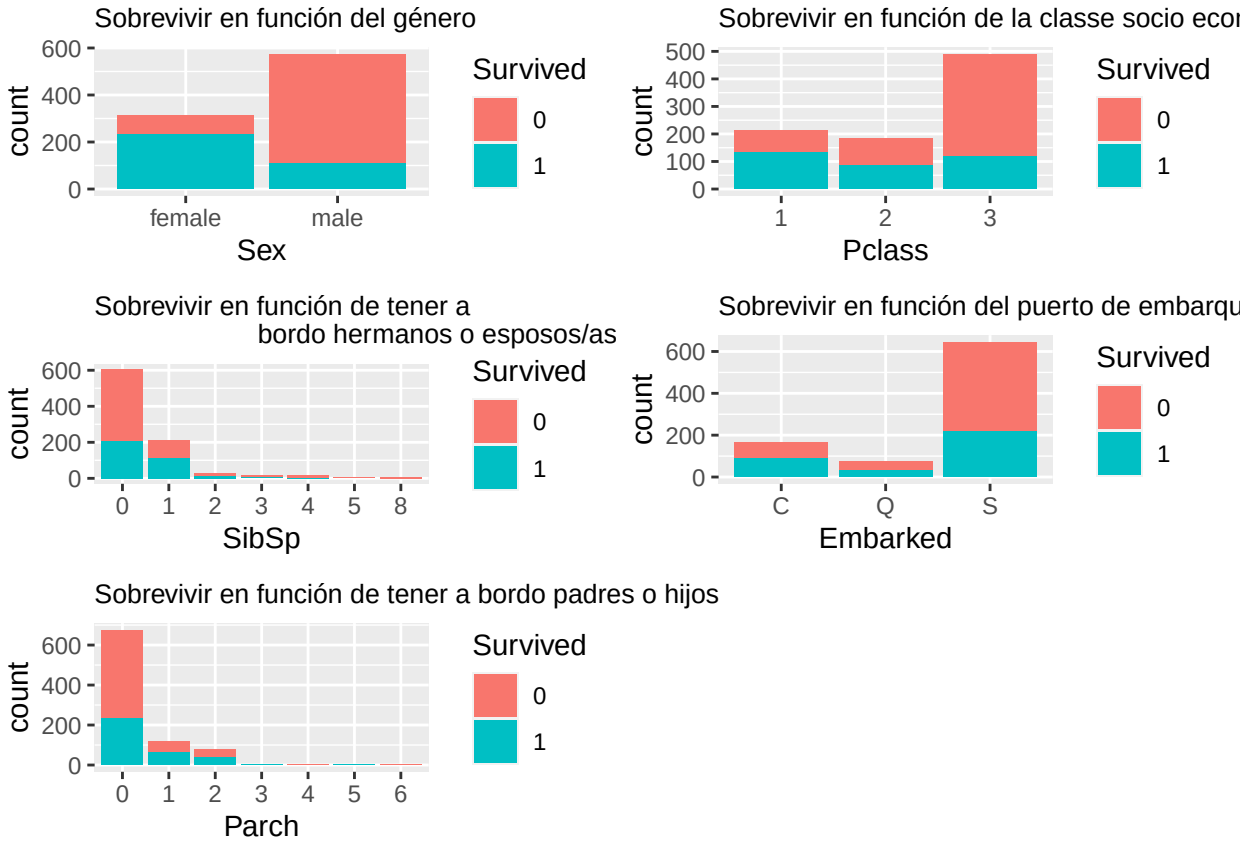
```
##
##  Pearson's Chi-squared test
##
## data:  Survived and Embarked
## X-squared = 23.755, df = 2, p-value = 6.944e-06
```

Hay relación.

4.3.1.2 Análisis visual Se analiza de forma visual las posibles correlaciones entre las variables categóricas del conjunto **Sex**, **Pclass**, **SibSp**, **Parch** y **Embarked** con **Survive**. Para ello, se utiliza un gráfico de frecuencias, que permite al mismo tiempo conocer mejor como se distribuyen estas variables.

```
# Histogramas
d.Sex <- ggplot(data = titanic, aes(x=Sex, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función del género") +
              theme(plot.title = element_text(size = 10))

d.Pclass <- ggplot(data = titanic, aes(x=Pclass, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función de la classe socio económica") +
              theme(plot.title = element_text(size = 10))

d.SibSp <- ggplot(data = titanic, aes(x=SibSp, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función de tener a
                      bordo hermanos o esposos/as") +
              theme(plot.title = element_text(size = 10))

d.Parch <- ggplot(data = titanic, aes(x=Parch, fill=Survived))+geom_bar()+
              ggtitle("Sobrevivir en función de tener a bordo padres o hijos") +
              theme(plot.title = element_text(size = 10))

d.Embarked <- ggplot(data = titanic, aes(x=Embarked, fill=Survived))+geom_bar()+
                ggtitle("Sobrevivir en función del puerto de embarque") +
                theme(plot.title = element_text(size = 10))
```

```
grid.arrange(d.Sex, d.Pclass, d.SibSp, d.Embarked, d.Parch,
             ncol = 2, nrow = 3)
```



Si se observa la variable **Sex**, hay mayor probabilidad de sobrevivir siendo mujer que siendo hombre.

Para el caso de de la variable que indica la clase socio económica, se observa una gran diferencia entre las probabilidades de sobrevivir en función de la clase. Más de la mitad de los pasajeros de primera clase sobrevivieron, de los pasajeros de segunda clase sobrevivieron aproximadamente la mitad, y para los de clase baja solo un cuarto lo hicieron, siendo esta la clase más numerosa.

Sobre las variables relacionadas con tener familia a bordo, **SibSp** y **Parch**, primero se destaca que la mayoría de pasajeros no tenían familiares a bordo, de ellos un poco más de un cuarto sobrevivieron. La segunda categoría más numerosa es tener un hermano o esposa/o a bordo, para el caso de **SibSp**, o tener un hijo o padre a bordo, para **Parch**, en ambos casos se observa que aproximadamente la mitad sobrevivieron, por lo que tener un familiar a bordo parece que influye en la probabilidad de sobrevivir. El mismo caso sucede para tener dos padres o hijos a bordo, aproximadamente la mitad sobreviven. Los demás casos donde hay más de dos familiares a bordo tienen tan pocas observaciones que no se pueden considerar relevantes para extraer conclusiones.

Finalmente, la variable **Embarked** nos indica en que puerto subieron los pasajeros. Se observa que la mayoría de pasajeros subieron en el puerto indicado como S, de los cuales menos de la mitad sobrevivieron. En el caso de pasajeros que subieron en los puertos indicados como C y Q, aproximadamente la mitad sobrevivieron.

Se puede intuir una relación directa con todas las variables categóricas observadas, y la probabilidad de sobrevivir.

4.3.1.3 Odds-Ratio A continuación, para representar de forma numérica las probabilidades de sobrevivir en función de la categoría de cada variable categórica, se calcularán los Odds-Ratio para la variable dependiente **Survived**. El OR es un número que permite conocer cómo varía la probabilidad de que la variable dependiente adquiera un valor en función del valor que adquiera la variable independiente.

Sex:

```
# Se calculan los diferentes Odds-Ratios

# Tabla de frecuencias
a <- as.data.frame(table(titanic$Survived, titanic$Sex))

odds.female <- a[2,]$Freq  / a[1,]$Freq
odds.male <- a[4,]$Freq  / a[3,]$Freq

cat("The OR between the female and the male was", odds.female / odds.male, "\n")
```

```
## The OR between the female and the male was 12.52752
```

Una mujer tenía 12,5 veces más probabilidades de sobrevivir que un hombre.

Pclass:

```
# Se calculan los diferentes Odds-Ratios

# Tabla de frecuencias
a <- as.data.frame(table(titanic$Survived, titanic$Pclass))

odds.high.class <- a[2,]$Freq  / a[1,]$Freq
odds.medium.class <- a[4,]$Freq  / a[3,]$Freq
odds.low.class <- a[6,]$Freq  / a[5,]$Freq

cat("The OR between the high class and the low class was", odds.high.class /
                                                          odds.low.class, "\n")
```

```
## The OR between the high class and the low class was 5.197059
```

```
cat("The OR between the medium class and the low class was", odds.medium.class /
                                                            odds.low.class, "\n")
```

```
## The OR between the medium class and the low class was 2.803777
```

```
cat("The OR between the high class and the medium class was", odds.high.class /
                                                             odds.medium.class, "\n")
```

```
## The OR between the high class and the medium class was 1.853592
```

Analizando los resultados, una persona de clase alta tenía 5,19 veces más probabilidades de sobrevivir que una persona de clase baja.

Una persona de clase media tenía casi el triple de probabilidades de sobrevivir que una persona de clase baja.

Una persona de clase alta tenía casi el doble de probabilidades de sobrevivir que una persona de clase media.

SibSp:

Para esta variable no se tendrán en cuenta las categorías con muy pocas observaciones.

```
# Tabla de frecuencias
a <- as.data.frame(table(titanic$Survived, titanic$SibSp))

odds.zero.SibSp <- a[2,]$Freq  / a[1,]$Freq
odds.one.SibSp <- a[4,]$Freq  / a[3,]$Freq
odds.two.SibSp <- a[6,]$Freq  / a[5,]$Freq

cat("The OR between two SibSp and zero SibSp", odds.two.SibSp /
                                               odds.zero.SibSp, "\n")
```

```
## The OR between two SibSp and zero SibSp 1.666345
```

```
cat("The OR between one SibSp and two SibSp", odds.one.SibSp /
                                              odds.two.SibSp, "\n")
```

```
## The OR between one SibSp and two SibSp 1.332276
```

```
cat("The OR between one SibSp and zero SibSp", odds.one.SibSp /
                                               odds.zero.SibSp, "\n")
```

```
## The OR between one SibSp and zero SibSp 2.220031
```

Observando los resultados, una persona con dos hermanos y/o esposos a bordo, tenia 1.6 veces más probabilidades de sobrevivir que una persona sin hermanos o esposos a bordo.

Una persona con un hermano o esposo, tenia 1.33 veces más probabilidades de sobrevivir que una persona con dos hermanos o esposos a bordo.

Una persona con un hermano o esposo a bordo, tenia 2.22 veces más probabilidades de sobrevivir que un pasajero sin hermanos o esposos a bordo.

Parch:

Para esta variable no se tendrán en cuenta las categorías con muy pocas observaciones.

```
# Tabla de frecuencias
a <- as.data.frame(table(titanic$Survived, titanic$Parch))

odds.zero.Parch <- a[2,]$Freq  / a[1,]$Freq
odds.one.Parch <- a[4,]$Freq  / a[3,]$Freq
odds.two.Parch <- a[6,]$Freq  / a[5,]$Freq

cat("The OR between two Parch  and zero Parch ", odds.two.Parch  /
      odds.zero.Parch , "\n")
```

```
## The OR between two Parch  and zero Parch  1.926407
```

```
cat("The OR between one Parch  and two Parch ", odds.one.Parch  /
      odds.two.Parch , "\n")
```

```
## The OR between one Parch  and two Parch  1.207547
```

```
cat("The OR between one Parch  and zero Parch ", odds.one.Parch  /
      odds.zero.Parch , "\n")
```

```
## The OR between one Parch  and zero Parch  2.326227
```

Observando los resultados, una persona con dos padres o hijos a bordo, tenia 1.93 veces más probabilidades de sobrevivir que una persona sin padres o hijos a bordo.

Una persona con un padre o hijo, tenia 1.21 veces más probabilidades de sobrevivir que una persona con dos padres o hijos a bordo.

Una persona con un padre o hijo a bordo, tenia 2.33 veces más probabilidades de sobrevivir que un pasajero sin padres o hijos a bordo.

Embarked:

```
# Tabla de frecuencias
a <- as.data.frame(table(titanic$Survived, titanic$Embarked))

odds.C <- a[2,]$Freq  / a[1,]$Freq
odds.Q <- a[4,]$Freq  / a[3,]$Freq
odds.S <- a[6,]$Freq  / a[5,]$Freq

cat("The OR between embarking in Q and embarking in S ", odds.Q  /
      odds.S , "\n")
```

```
## The OR between embarking in Q and embarking in S  1.88
```

```
cat("The OR between embarking in C and embarking in Q ", odds.C  /
      odds.Q , "\n")
```

```
## The OR between embarking in C and embarking in Q  NaN
```

```
cat("The OR between embarking in C and embarking in S ", odds.C  /
      odds.S , "\n")
```

```
## The OR between embarking in C and embarking in S  NaN
```

Se observa que una persona que hubiera embarcado en el puerto Q, tenia 1.24 veces más probabilidades de sobrevivir que una persona que hubiera embarcado en el puerto S.

Una persona que hubiera embarcado en el puerto C, tenia 1.88 veces más probabilidades de sobrevivir que una persona que hubiera embarcado en el puerto Q.

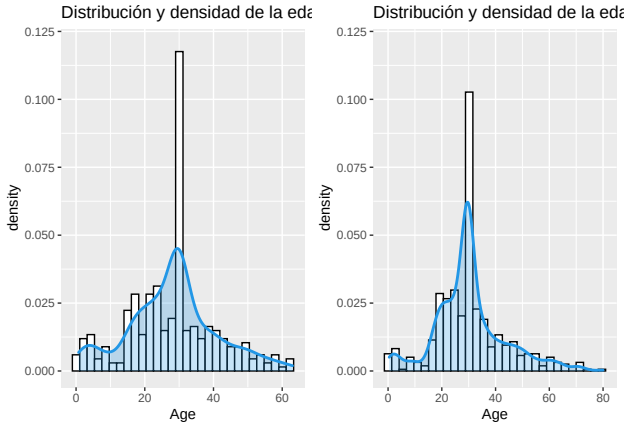
Finalmente, una persona que hubiera embarcado en el puerto C, tenia 2.34 veces más probabilidades de sobrevivir, que una persona que hubiera embarcado en el puerto S.

4.3.2 ¿La media de edad de las mujeres es igual o inferior a la de los hombres?

El objetivo es comprobar si la media de edad de las mujeres es igual o inferior a la de los hombres. Para responder a esta pregunta, se realiza una prueba estadística por contraste de hipótesis.

4.3.2.1 Análisis visual Se observan las distribuciones y densidades de las edades en función del genero.

```
p1 <- ggplot(titanic.women, aes(x = Age)) +
        geom_histogram(aes(y = ..density..), bins = 30,
                       colour = 1, fill = "white") +
        geom_density(lwd = 1, colour = 4,
                     fill = 4, alpha = 0.25) +
        scale_y_continuous(limit = c(0,0.12)) +
        ggtitle("Distribución y densidad de la edad de las mujeres")

p2 <- ggplot(titanic.men, aes(x = Age)) +
        geom_histogram(aes(y = ..density..), bins = 30,
                       colour = 1, fill = "white") +
        geom_density(lwd = 1, colour = 4,
                     fill = 4, alpha = 0.25) +
        scale_y_continuous(limit = c(0,0.12)) +
        ggtitle("Distribución y densidad de la edad de los hombres")

grid.arrange(p1,p2,ncol = 2, nrow = 1)
```



En ambas distribuciones se observa un pico pronunciado en los 30 años. La edad de las mujeres va de cero a sesenta y pocos años mientras que la de los hombres va de cero a ochenta. Las distribuciones son parecidas, pero no parecen muy normales, pues no tienen forma simétrica de campana de Gauss.

4.3.2.2 Normalidad y homoscedasticidad Se realiza una prueba de normalidad Shapiro-Wilk para determinar si se trata de distribuciones normales.

Women:

```
shapiro.test(titanic.women$Age)
```

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

```
##
## data:  titanic.women$Age
## W = 0.97511, p-value = 3.009e-05
```

Men:

```
shapiro.test(titanic.men$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  titanic.men$Age
## W = 0.94348, p-value = 5.472e-14
```

Se recuerda que para los tests de normalidad, la hipótesis nula indica normalidad, mientras que la alternativa indica que las distribuciones no son normales. En ambos casos, se obtienen valores p inferiores a los límites comunes de significancia, por ello, se confirma que son distribuciones normales.

También es relevante mencionar que al disponer de una muestra poblacional suficientemente grande, según el Teorema del Límite Central, la distribución de las medias muéstrales seguirán una distribución normal y por ello se podrá aplicar la posterior prueba por contraste de hipótesis.

A continuación, se evalúa la homoscedasticidad entre las edades de los hombres y las mujeres, para determinar si tienen varianzas iguales o diferentes. La hipótesis nula de este test indica que las varianzas son iguales.

```
var.test(titanic.women$Age, titanic.men$Age)
```

```
##
##  F test to compare two variances
##
## data:  titanic.women$Age and titanic.men$Age
## F = 0.97918, num df = 312, denom df = 574, p-value = 0.8404
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.807806 1.193718
## sample estimates:
## ratio of variances
##           0.979184
```

Como el p-valor es superior a los límites comunes de significancia, no se puede rechazar la hipótesis nula de que las varianzas son iguales.

4.3.2.3 Contraste de hipótesis Para realizar un estudio por contraste de hipótesis, primeramente se plantea la pregunta responder, la hipótesis nula y la hipótesis alternativa.

Pregunta: La media de edad de las mujeres es igual o inferior a la de los hombres?

Hipótesis nula: $H_0 : \mu_W = \mu_M$

Hipótesis alternativa: $H_1 : \mu_W < \mu_M$

Las variables se pueden considerar las distribuciones normales al aplicar el Teorema del límite Central, y que las varianzas de ambas muestras son iguales. Se aplica entonces un test de hipótesis paramétrico de dos muestras independientes sobre la media de edad. Se tratará de un test unilateral por la izquierda, es decir, para rechazar la hipótesis nula, se debe observar una media de edad suficientemente inferior para las mujeres versus la media de edad de los hombres.

```r
t.test(titanic.women$Age, titanic.men$Age, alternative = "l", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  titanic.women$Age and titanic.men$Age
## t = -2.5152, df = 886, p-value = 0.006036
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.7920278
## sample estimates:
## mean of x mean of y
##  28.19506  30.48845
```

Si se considera un nivel de significancia de $\alpha = 0.05$, se observa un valor p inferior a α, por lo que se puede descartar la hipótesis nula que indica que ambas medias son iguales, y nos decantaremos por la hipótesis alternativa que indica que la media de edad de las mujeres es inferior a la de los hombres. Los resultados del test indican que la media muestral de las mujeres es de 28.19 mientras que la de los hombres es de 30.49 años de edad.

4.4 Modelo de regresión logística

En este apartado, se crea un modelo de regresión logística, y se analiza como añadir variables independientes al modelo afecta la precisión de los resultados (*accuracy*).

En primer lugar, se divide el dataset en un dataset de train y otro de test que se usará para evaluar los resultados de los modelos. Se realiza esta partición manteniendo una proporción 3 a 1.

```r
set.seed(23)

trainIndex=createDataPartition(titanic$Survived, p=0.75)$Resample1

titanic_train=titanic[trainIndex, ]
titanic_test= titanic[-trainIndex, ]

cat("The train set has", nrow(titanic_train), "rows\n\n")
```

```
## The train set has 667 rows
```

```r
cat("The train set has", nrow(titanic_test), "rows")
```

```
## The train set has 221 rows
```

Para evaluar los modelos, se calcula su precisión a partir del dataset de test.

4.4.1 Modelo base En primer lugar, se crea un modelo que utiliza una solo variable para predecir la variable dependiente **Survived** y se evalúa su precisión. Este proceso ayudará a determinar las variables más relevantes para la predicción.

```
for (var in colnames(titanic_train)){
  if (var != "Survived"){
    formula <- as.formula(sprintf("Survived ~ %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with ", var, " is ", accuracy, "\n")
  }
}
```

```
## The accuracy of the model with  Pclass  is  0.7104072
## The accuracy of the model with  Sex  is  0.7692308
## The accuracy of the model with  Age  is  0.6199095
## The accuracy of the model with  SibSp  is  0.6742081
## The accuracy of the model with  Parch  is  0.5972851
## The accuracy of the model with  Fare  is  0.6968326
## The accuracy of the model with  Embarked  is  0.6334842
```

Se observa que la variable con la que obtenemos un *accuracy* mayor es **Sex**, como ya se estudió previamente.

A continuación, se irán añadiendo el resto de variables al modelo con el objetivo de mejorar las predicciones.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived"){
    formula <- as.formula(sprintf("Survived ~ Sex * %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex and ", var, " is ", accuracy, "\n")
  }
}
```

4.4.2 Sex y el resto de variables

```
## The accuracy of the model with Sex and  Pclass  is  0.7692308
## The accuracy of the model with Sex and  Age  is  0.7692308
## The accuracy of the model with Sex and  SibSp  is  0.7782805
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
## The accuracy of the model with Sex and  Parch  is  0.7737557
## The accuracy of the model with Sex and  Fare  is  0.7692308
## The accuracy of the model with Sex and  Embarked  is  0.7692308
```

La variable que mejores resultados muestra es **SibSp**. Además, la variable **Parch** muestra un warning de correlación con la variable **Sex**, por lo que se decide incluir la variable **SibSp** en el modelo.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived" && var != "SibSp" && var != "Parch"){
    formula <- as.formula(sprintf("Survived ~ Sex * SibSp + %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex, SibSp and ", var, " is ", accuracy,
        "\n")
  }
}
```

4.4.3 Sex, SibSp y el resto de variables

```
## The accuracy of the model with Sex, SibSp and  Pclass  is  0.7828054
## The accuracy of the model with Sex, SibSp and  Age  is  0.7782805
## The accuracy of the model with Sex, SibSp and  Fare  is  0.7782805
## The accuracy of the model with Sex, SibSp and  Embarked  is  0.7782805
```

Se hace el estudio sin las interacciones, ya que la gran cantidad de términos de correlación harían que los modelos generados no fuesen fiables.

Se decide añadir ahora la variable **Pclass**, siendo la variable que mejores resultados proporciona.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived" && var != "SibSp" && var != "Parch"
      && var != "Pclass"){
    formula <- as.formula(sprintf("Survived ~ Sex * SibSp + Pclass * %s", var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex, SibSp, Pclass and ", var, " is ",
        accuracy, "\n")
  }
}
```

4.4.4 Sex, SibSp, Pclass y el resto de variables.

```
## The accuracy of the model with Sex, SibSp, Pclass and  Age  is  0.8099548
## The accuracy of the model with Sex, SibSp, Pclass and  Fare  is  0.7828054
## The accuracy of the model with Sex, SibSp, Pclass and  Embarked  is  0.7828054
```

Se observa que al añadir al modelo la variable **Age**, se consigue un *accuracy* de 0.81, siendo esta la variable que mejores resultados aporta. Se añade al modelo.

```
for (var in colnames(titanic_train)){
  if (var != "Sex" && var != "Survived" && var != "SibSp" && var != "Parch"
      && var != "Pclass" && var != "Age"){
    formula <- as.formula(sprintf("Survived ~ Sex * SibSp + Pclass * Age * %s",
                                  var))
    model <- glm(formula, family = binomial(link=logit), data=titanic_train)

    titanic_test$pred <- predict(model, titanic_test, type = "response")
    titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
    titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d,
                                 1, 0)
    accuracy <- sum(titanic_test$is_ok)/nrow(titanic_test)
    cat("The accuracy of the model with Sex, SibSp, Pclass, Age and ", var,
        " is ", accuracy, "\n")
  }
}
```

4.4.5 Sex, SibSp, Pclass, Age y el resto de variables.

```
## The accuracy of the model with Sex, SibSp, Pclass, Age and  Fare  is  0.7963801
## The accuracy of the model with Sex, SibSp, Pclass, Age and  Embarked  is  0.7918552
```

Se observa que llegado este punto, la precisión del modelo al añadir las variables restantes **Fare** y **Embarked** disminuye. Por esta razón, se determina que el modelo final es el que incluye las variables, **Sex, SibSp, Pclass y Age**.

```
model.final <- glm(Survived ~ Sex * SibSp + Pclass * Age,
                   family = binomial(link=logit), data=titanic_train)
titanic_test$pred <- predict(model.final, titanic_test, type = "response")
titanic_test$pred_d <- ifelse(titanic_test$pred < 0.5, 0, 1)
titanic_test$is_ok <- ifelse(titanic_test$Survived == titanic_test$pred_d, 1, 0)
accuracy.final <- sum(titanic_test$is_ok)/nrow(titanic_test)
cat("The accuracy of the final model is ", accuracy.final)
```

4.4.6 Modelo final

```
## The accuracy of the final model is  0.8099548
```

La precisión del modelo final es de 0.81, es decir, se han predicho correctamente el 81% de las observaciones del conjunto de datos test.

Ejercicio 5

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

Con la realización de esta práctica, se ha querido responder a la pregunta, **"¿Cuáles fueron los factores de peso que influyeron en que los pasajeros del Titanic tuvieran mayor probabilidad de sobrevivir?"**

Se partió del conjunto de datos del Titanic proporcionado por los recursos de la práctica, se seleccionaron las variables que describían características de los pasajeros y se realizó un proceso de limpieza y tratamiento de valores vacíos y outliers.

Seguidamente, se realizó un estudio de normalidad y homogeniedad de la varianza para las variables numéricas del conjunto de datos **Age** y **Fare**, del cual se pudo determinar que ninguna de ellas sigue una distribución normal, ni las varianzas de **Age** y **Fare**, entre las dos categorías de la variable dependiente estudiada **Survived** son iguales. Es importante mencionar, que el conjunto de datos cuenta con 888 observaciones, por lo que se puede aplicar el Teorema del Límite Central, que indica que dada una muestra de población suficientemente grande, la distribución de las medias muéstrales seguirá una distribución normal.

A continuación, se realizó un estudio sobre la relación entre la variable dependiente **Survived** y las variables categóricas del conjunto de datos. Mediante la prueba Chi-cuadrado, se determinó que todas las variables categóricas tienen relación con **Survived**. Del análisis visual y el cálculo del Odds-Ratio, se concluyó que los factores que pueden tener más relevancia sobre la probabilidad de sobrevivir al accidente son **Sex** y **Pclass**, siendo estos los factores que más variabilidad presentan en la probabilidad de sobrevivir entre sus distintas categorías.

Para profundizar sobre la variable numérica **Age**, se planteó la pregunta ¿La media de edad de las mujeres es igual o inferior a la de los hombres? Para responder a esta pregunta, se realzió un estudio por contraste de hipótesis y se determinó que con un nivel de significancia del 0.05% que se puede confirmar que la media de edad de las mujeres es inferior a la de los hombres.

Finalmente, se elaboró un modelo de regresión logística con el objetivo de predecir si un pasajero sobrevivió o no en función de distintos factores. Se dividió el conjunto de datos en un subset de train y otro de test para poder evaluar la precisión de los resultados. Se destacan las siguientes conclusiones:

En primer lugar, la variable más determinante a la hora de discernir entre si un pasajero sobrevivió o no fue la variable **Sex**, un resultado esperado en vista de los análisis previamente realizados. Sin embargo, la segunda variable que se incluyó en el modelo, por su significancia demostrada al mejorar los resultados de las predicciones, fue la variable **SibSp**, lo cual era difícilmente predecible en función de los estudios previos. En cada paso posterior, se añadió al modelo la variable que mejor precisión otorgaba al ser añadida, hasta alcanzar el punto en que no se observó mejora al seguir añadiendo variables. El modelo final estaba compuesto por las variables **Sex**, **SibSp**, **Pclass** y **Age** por lo que se considera que estas cuatro variables son las **más influyentes en la probabilidad de sobrevivir**. La precisión final del modelo es del 81%.

Se destaca que varias variables se añadieron al modelo sin los términos de interacción, ya que el escueto tamaño del dataset no permitía la creación de modelos con tantos términos.

Tabla de contribuciones

```
##                 Contribuciones                                  Firma
## 1         Investigación previa Marta Coll Pol, Manuel De Blas Pino
## 2 Redacción de las respuestas  Marta Coll Pol, Manuel De Blas Pino
## 3           Desarrollo código  Marta Coll Pol, Manuel De Blas Pino
```

Referencias

[1] Histogram with density in ggplot2 [fecha de consulta: 30/05/22] Disponible en: https://r-charts.com/distribution/histogram-density-ggplot2/

[2] Sigüeñas Gonzales, Manuel. Pruebas de Normalidad [fecha de publicación: 28/10/2015] [fecha de consulta: 30/05/22] Disponible en: https://rpubs.com/MSiguenas/122473

[3] Amat Rodrigo, Joaquín. Análisis de la homogeneidad de varianza (homocedasticidad)[fecha de publicación: 01/01/2016] [fecha de consulta: 30/05/22] Disponible en: https://rpubs.com/Joaquin_AR/218466

[4] The Chi-Square Test for Independence [fecha de consulta: 30/05/22] Disponible en: https://soc.utah.edu/sociology3112/chi-square.php

[5] Warning in R - Chi-squared approximation may be incorrect [fecha de consulta: 30/05/22] Disponible en: https://stats.stackexchange.com/questions/81483/warning-in-r-chi-squared-approximation-may-be-incorrect