Titanic PRA2: Limpieza y análisis de datos

Autores

María Ángeles Fuentes Expósito

Norberto Jesús de la Cruz Falcón

NOTA: ESTE DOCUMENTO SINTETIZA LOS RESULTADOS DEL ANÁLISIS OBTENIDOS EN RSTUDIO, PUEDE CONSULTAR EL CÓDIGO Y LA EJECUCIÓN COMPLETA EN PRAC2.RMD

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

Estado del arte

El Titanic era un barco de 267 metros de eslora que choca contra un iceberg y comienza a hundirse. Los pasajeros, aterrorizados, esperan su turno para subir a uno de los pocos botes salvavidas que hay: 20 en total. De las 2.223 personas que viajaban en él, solo 706 (492 pasajeros y 214 tripulantes) sobrevivieron tras ser recogidas a la mañana siguiente. Del agua se recuperaron 333 cuerpos.

No fue el naufragio más grande de la historia en número de víctimas –en 1945, por ejemplo, murieron 9.343 personas en el accidente del Wilhelm Gustloff, un transatlántico alemán hundido por un submarino soviético–, pero el hundimiento del Titanic sí fue la catástrofe más mítica por la notoriedad de algunas de las víctimas (de las más ricas del mundo), por tratarse

del viaje inaugural de un barco de superlujo y por los enigmas que, a día de hoy, siguen rodeando a este suceso [1].

Justificación

La pregunta que pretendemos responder es: ¿podría sobrevivir al hundimiento del Titanic? ¿existen características o patrones que tienen en común los supervivientes?, ¿se dio prioridad de acceso a los botes salvavidas mujeres y niños?, ¿hubo influencia económica (de clase) en los supervivientes?

Descripción del dataset

Partimos de una base de datos real con los nombres y apellidos de una muestra de los pasajeros del Titanic (891 registrados frente a los 2.000 que viajaron), con su edad, el número de familiares de cada uno, la clase en que viajaban y el precio del billete que habían pagado. También sabemos si sobrevivieron o no. El dataset se obtiene de la competición de *Kaggle "Titanic - Machine Learning from Disaster*" [2] disponible en el repositorio y está formado por las siguientes variables descritas en la Tabla 1.

Tabla 1 Tabla descripción atributos dataset

Variable	Definición	Clave
PassengerId	Número identificador del pasajero	
Survived	Variable dicotómica que indica si el pasajero sobrevivió al naufragio	0 = No, 1 = Yes
Pclass	Ticked de Clase a la que pertenecía el pasajero	1 = 1st, 2 = 2nd, 3 = 3rd
Name	Nombre y apellidos del pasajero	
Sex	Género del pasajero	
Age	Edad del pasajero	
SibSp	Número de hermanos o de cónyuges que tenía el pasajero a bordo	
Parch	Número de hijos o padres que tenía el pasajero a bordo	
Ticket	Número identificador del ticket	
Fare	Tarifa del billete del pasajero en libras esterlinas	
Cabin	Cabina asignada al pasajero	
Embarked	Puerto en el que embarcó el pasajero	C = Cherbourg, Q = Queenstown, S = Southampton

2. Integración y selección de datos de interés.

Este proceso requiere visualizar el conjunto de datos, comprenderlos y seleccionar los atributos que son necesarios para realizar el estudio. Por lo que describimos a continuación los pasos realizados:

1. Cargamos el conjunto de datos y observamos que está formado por 891 registros y las 12 variables comentadas en el apartado anterior (Tabla 1).
2. Generaremos un dataset que será la subselección útil de los datos originales, ya que las variables *Name* y *Ticket* no serán de utilidad en nuestro análisis.

```
head(df)
```

```
##   PassengerId Survived Pclass    Sex Age SibSp Parch    Fare Cabin Embarked
## 1           1        0      3   male  22     1     0  7.2500              S
## 2           2        1      1 female  38     1     0 71.2833   C85        C
## 3           3        1      3 female  26     0     0  7.9250              S
## 4           4        1      1 female  35     1     0 53.1000  C123        S
## 5           5        0      3   male  35     0     0  8.0500              S
## 6           6        0      3   male  NA     0     0  8.4583              Q
```

Ilustración 1 Subselección datos

3. Lo siguiente que hacemos es verificar el tipo de dato de los atributos y observamos que los atributos cualitativos *Pclass* , *Sex* y *Cabin* no se han cargado en Rstudio como tipo **Factor** `class(df$Pclass), class(df$Sex),class(df$Cabin)`, por lo que se realiza la conversión al tipo de dato Factor.

En resumen, se han seleccionado y configurado los atributos de interés del dataset de supervivientes de Titanic y realizado la primera vista de los registros.

3. Limpieza de los datos.

3.1. Valores perdidos.

Una de las técnicas de limpieza de los datos es la eliminación de los valores nulos o *missing data*, es importante analizar y estudiar cómo realizar el tratamiento de los valores perdidos, ya que si eliminamos los registros podríamos perder información relevante y si añadimos datos a los valores perdidos podríamos crear sesgos e información falsa en los datos.

Lo primero que se ha de hacer es visualizar el conjunto de datos y capturar los valores nulos NA's.

```
summary(df)
```

```
##   PassengerId       Survived      Pclass      Sex           Age       
##  Min.   :  1.0   Min.   :0.0000   1:216   female:314   Min.   : 0.42  
##  1st Qu.:223.5   1st Qu.:0.0000   2:184   male  :577   1st Qu.:20.12  
##  Median :446.0   Median :0.0000   3:491                Median :28.00  
##  Mean   :446.0   Mean   :0.3838                        Mean   :29.70  
##  3rd Qu.:668.5   3rd Qu.:1.0000                        3rd Qu.:38.00  
##  Max.   :891.0   Max.   :1.0000                        Max.   :80.00  
##                                                        NA's   :177    
##      SibSp           Parch             Fare                Cabin     
##  Min.   :0.000   Min.   :0.0000   Min.   :  0.00              :687  
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:  7.91   B96 B98    :  4  
##  Median :0.000   Median :0.0000   Median : 14.45   C23 C25 C27:  4  
##  Mean   :0.523   Mean   :0.3816   Mean   : 32.20   G6         :  4  
##  3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.: 31.00   C22 C26    :  3  
##  Max.   :8.000   Max.   :6.0000   Max.   :512.33   D          :  3  
##                                                    (Other)    :186  
##    Embarked
##  Length:891
##  Class :character
##  Mode  :character
##
```

Ilustración 2 Resumen del conjunto de datos

Observamos que la variable *Age* tiene 177 valores nulos y si eliminamos estos registros perdemos un 10% de los datos del conjunto. Claramente tenemos que proceder a evaluarlos para realizar la imputación de estos registros.

Para lograr una imputación lo más acertada posible, analizaremos si existe alguna relación en la pérdida de datos con otros factores, como PClass y Age.

```
sum(is.na(df$Age))
```

```
## [1] 177
```

```
sum(is.na(df[df$Pclass == 1,]$Age))
```

```
## [1] 30
```

```
sum(is.na(df[df$Pclass == 2,]$Age))
```

```
## [1] 11
```

```
sum(is.na(df[df$Pclass == 3,]$Age))
```

```
## [1] 136
```

Ilustración 3 Valores perdidos por cada Clase

Se observa que existen 177 valores perdidos en las edades de los pasajeros del Titanic, y que además la mayor parte (136) pertenece a los pasajeros de tercera clase. Este dato puede ser interesante para el estudio y conclusiones del análisis.

Procedemos a agrupar los registros perdidos en la tercera clase para ver si existe relación con el género de los pasajeros.

En la Ilustración 4, observamos que los *missing values* de *Age* se encuentran mayoritariamente en la Tercera Clase y además mayoritariamente en hombres.

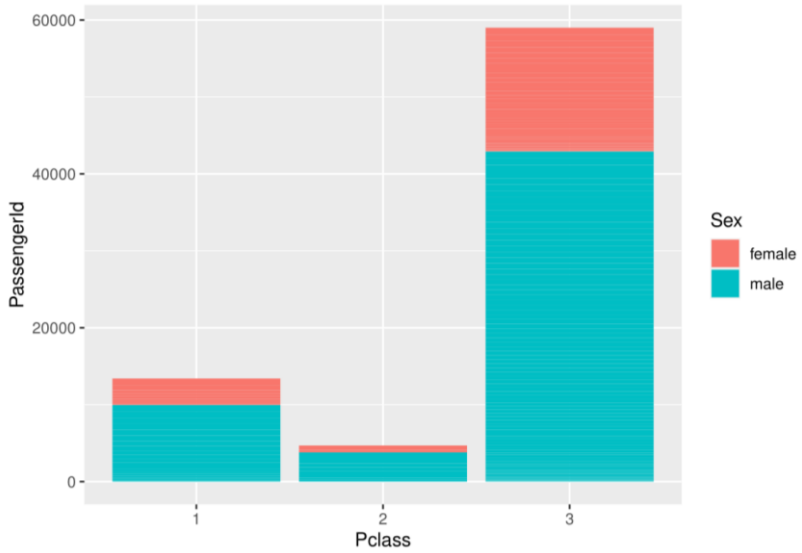

Ilustración 4 Pasajeros que no tienen valor en el registro Age según Clase y Sexo

Procedemos a realizar la Imputación de los registros de edad mediante la técnica de clustering kNN para cada Clase.

Ahora sí hemos imputado las edades para todos los pasajeros y no hemos de eliminar ningún registro.

Mediante el código en R `levels(as.factor(df$ATRIBUTO))` comprobamos si existen missing values en cada una de las variables.

- En la variable *Survived* no se observan valores pérdidos.
- Tampoco se observan valores pérdidos en la variable *Pclass*. La variable Pclass únicamente toma los valores 1, 2 y 3.
- La variable *Sex* toma sólo dos valores male y female y no existen valores pérdidos.

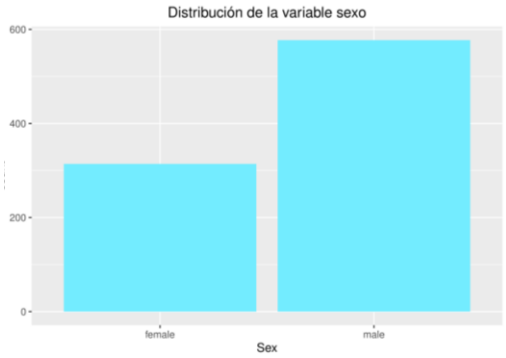

Ilustración 5 Distribución de la variable sexo

- Tampoco se observan valores pérdidos en la variable Parch ni en la variable SibSp.

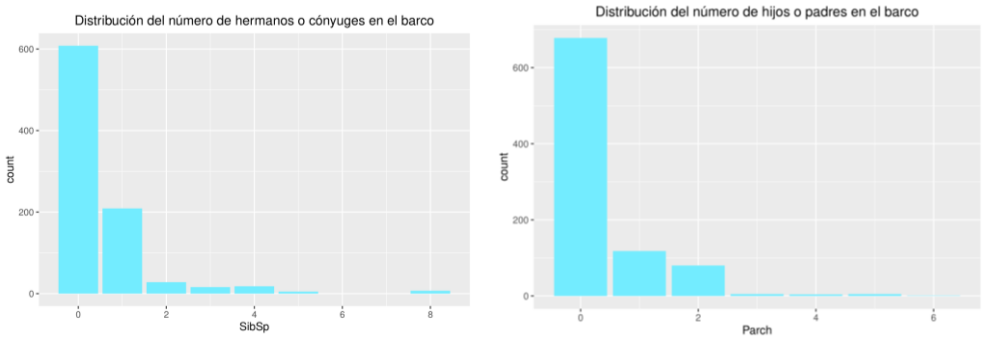

Ilustración 6 Distribución de la variable SibSp y de la variable Parch

- Por el contrario, existen valores vacíos en la variable *Cabin*. En concreto, más de la mitad de los registros tienen valores vacíos en la variable *Cabin* (529 de 891). Entre los pasajeros de segunda y tercera clase 600 inmigrantes embarcaron en busca de una oportunidad y viajaron en condiciones muy diferentes a los demás viajeros [3].

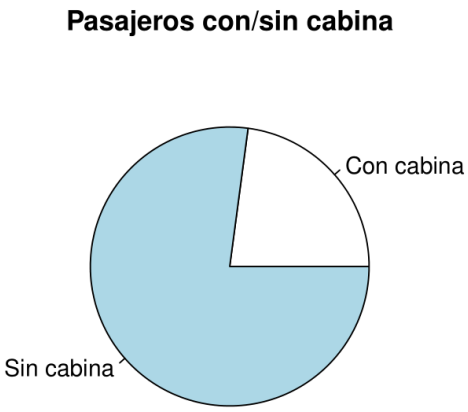

Ilustración 7 Pasajeros con o sin cabina

- Por último se encuentran dos registros con valores perdidos en la variable *Embarked*. Se eliminan estos dos registros.

Leyenda: C = Cherbourg, Q = Queenstown, S= Southampton

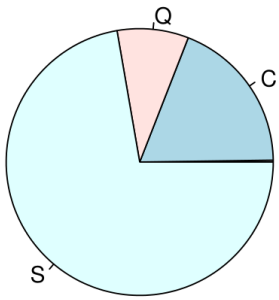

Ilustración 8 Pasajeros según el puerto de embarque

3.2. Valores extremos.

Los diagramas de bigotes nos dan una visual rápida sobre las distribución de los valores y los *outliers*. Vemos que la media de la edad ronda los 30 años, y que tenemos valores extremos entre los 60 y 80 años. No consideramos eliminarlos de la muestra, ya que pueden ser datos interesantes y no se consideran valores erróneos.

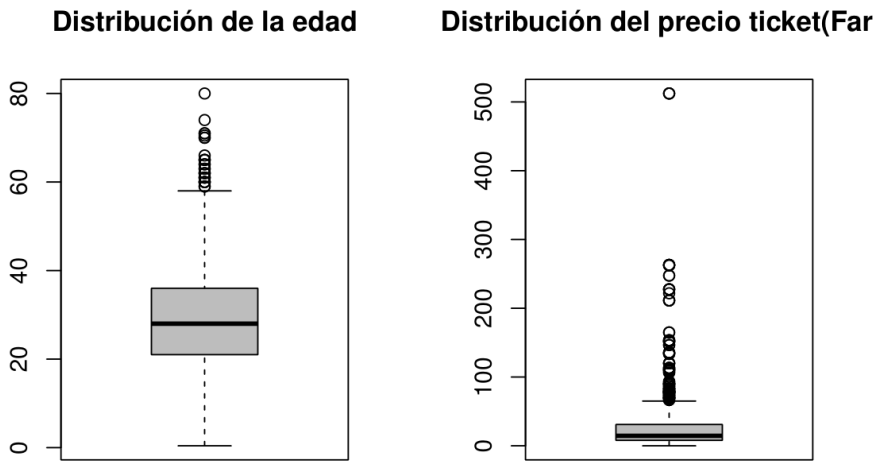

Ilustración 9 Gráfico boxplot de las variables age y fare

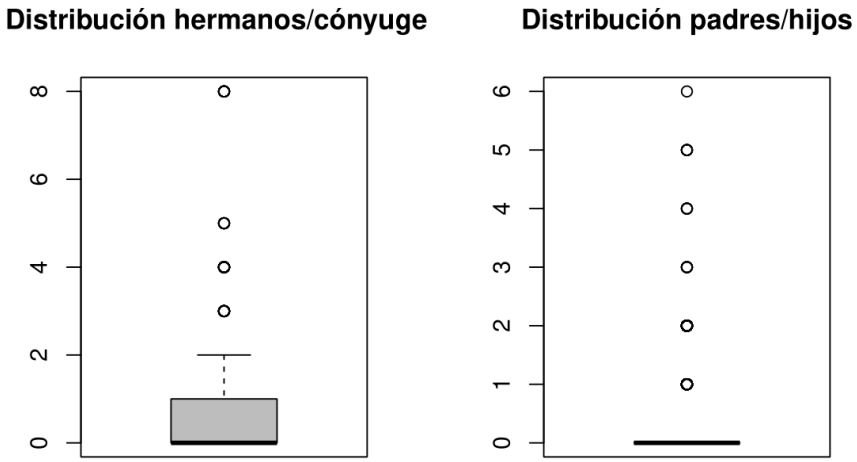

Ilustración 10 Gráficos boxplot de las variables sibSp y parch

Ocurre algo, sin embargo, con el precio de los tickets, el precio medio se encuentra entre 32 y aparece un ticket con precio 512. Además, vemos que hay tickets que costaron 0 libras esterlinas. Según nuestra investigación en la red: el pasaje oscilaba entre 3 libras, los más baratos, hasta 870 libras, los más caros [4].

```
#resumen valores de atributo Precio del ticket
summary (df$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.896  14.454  32.097  31.000 512.329
```

Ilustración 11 Valores min, media, cuartiles y máx encontrado en el conjunto de datos

Discutimos si se trata de un dato erróneo o no, y concluimos que no imputaremos estos valores extremos porque pueden ser muy valiosos a la hora de estudiar la razón y las relaciones que pudieran aparecer respecto al precio de los billetes.

A continuación, distribuimos la edad según la clase: primera clase, segunda clase y tercera clase.

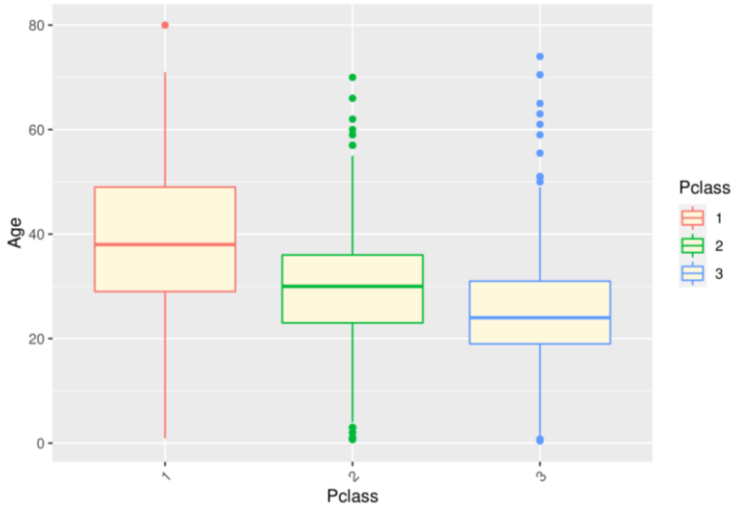

Ilustración 12 Gráfico boxplot de la variable age para cada una de las clases

Los valores extremos aparecen, aunque son los que se dibujan en los límites inferiores los que levantan sospecha ya que se trata de infantes. Los extremos superiores, a simple vista, determinan que la media de edad en la segunda y tercera clase es más joven que primera clase (30 años y 25 años frente a 38 años) y que son pocos los mayores de 55 años que viajaron en segunda y tercera clase.

Sobre el número de hermanos o de cónyuges que tenía el pasajero a bordo, vemos que el valor parece ser bastante irrelevante, la media es 0.5, y en tercera clase hay valores extremos de algunos pasajeros que viajaban con hermanos o cónyuges.

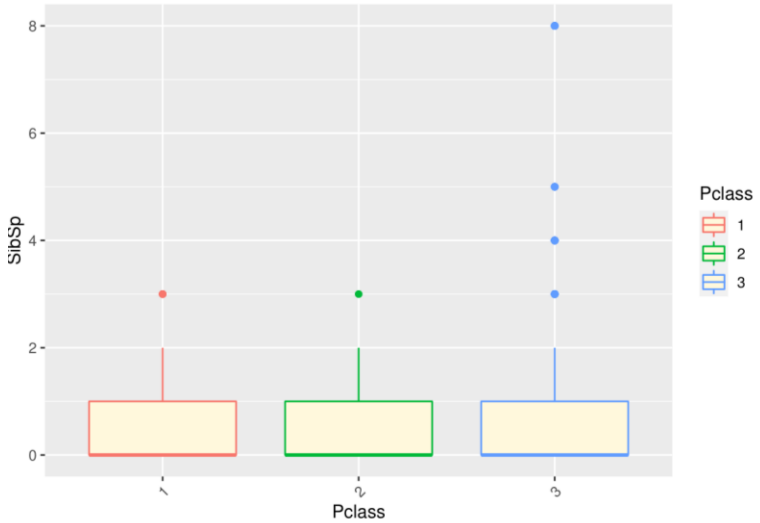

Ilustración 13 Gráfico boxplot de la variable sibsp para cada una de las clases

Sin embargo en el atributo *Parch* sobre el número de hijos o padres a bordo vemos que hay *outliers* en primera y tercera clase, pero que en segunda clase si se encuentran relaciones de padres o hijos a bordo.

```
#resumen valores de número de padres o hijos a bordo
summary (df$Parch)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3825  0.0000  6.0000
```

Ilustración 14 Parámetros estadísticos de la variable parch

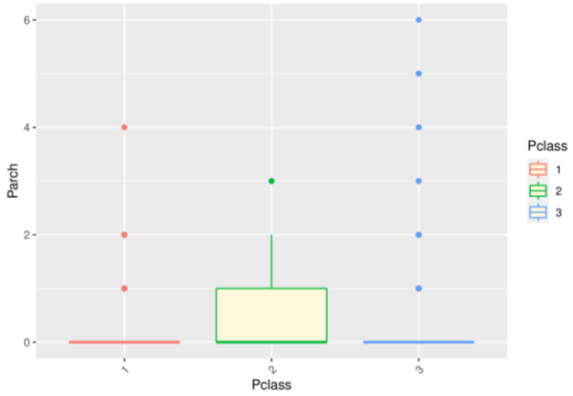

Ilustración 15 Gráfico boxplot de la variable parch para cada una de las clases

Decidimos no tratar los *outliers* o valores extremos, porque pueden ser datos interesantes a tener en cuenta durante el análisis.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar

Tras conocer los atributos y los valores que contiene el conjunto de datos de formas general, podemos definir los siguientes grupos (Tabla 2) para comparar y realizar los análisis que propuestos a continuación.

Tabla 2 Grupos seleccionados

Grupo	Atributo key
Mujeres y hombres	Sex (Gender)
Primera clase, segunda clase, tercera clase	Pclass
Superviviente y no superviviente	Survived
Niños, adultos, ancianos	Age range (generated)

Análisis y tipo de análisis a aplicar

Supervivencia entre mujeres y hombres.	→	regresión logística con variable dependiente *Survived* y variable explicativa *Sex*.
Supervivencia es mayor en los niños.	→	contraste de hipótesis sobre la proporción de supervivencia entre dos muestras (niños y adultos) atributo *Age*.
Existe relación de supervivencia en niños y clase.	→	contraste de correlación entre *Survived*, *Age* y *Pclass: regresión logistica.*
La supervivencia es mayor en viajeros de primera clase.	→	Regresión logística.
Existe relación entre el precio del billete y el puerto de embarque.	→	Test de chi-cuadrado discretizando el precio del billete por intervalos.
La supervivencia es mayor en pasajeros que tenían cabina asignada.	→	Regresión logística.
Las variables has_cabine y Pclass están relacionadas o son independientes.	→	Test de chi-cuadrado.

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

Cuando estudiamos la **normalidad** de los datos podemos comenzar utilizando el gráfico cuantil-cuantil (QQ Plot) para comprobar si los valores de nuestro conjunto se distribuyen normalmente. En nuestro caso, visualizamos a la izquierda la normalidad de la variable *Age,* en la que parece que presenta cierta tendencia a la normalidad, mientras que la gráfica de *Fare* claramente no sigue la forma de una distribución normal.

normalmente(QQ plot)

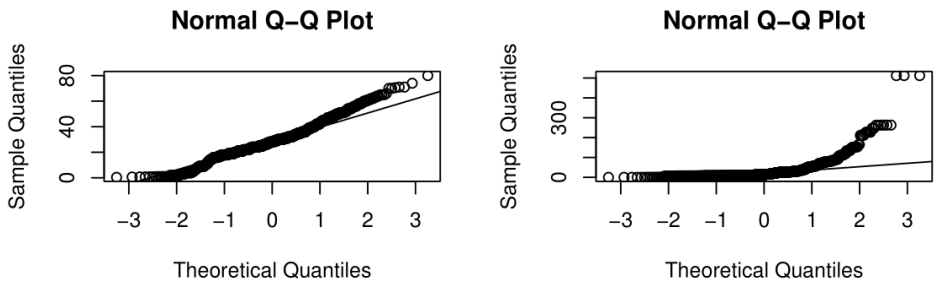

Ilustración 6 Gráfico cuantil-cuantil - Izquierda Age, derecha Fare

Otra forma potente para determinar si las variables *Age* y *Fare* siguen una distribución normal es emplear el **test de Shapiro-Wilk** [5]. Para el test de Shapiro-Wilk suponemos un nivel de significancia igual a 0.05.

Siendo la hipótesis nula que la población está distribuida normalmente, si el p-valor es menor al nivel de significancia entonces la hipótesis nula es rechazada (se concluye que los datos no vienen de una distribución normal). Si el p-value es mayor al valor de significancia, se concluye que no se puede rechazar dicha hipótesis.

```
shapiro.test(df$Age)

##
##  Shapiro-Wilk normality test
##
## data:  df$Age
## W = 0.97818, p-value = 2.868e-10
```

```
shapiro.test(df$Fare)

##
##  Shapiro-Wilk normality test
##
## data:  df$Fare
## W = 0.5197, p-value < 2.2e-16
```

Ilustración 16 Resultado del test Shapiro-Wilk para las variables age y fare

- Tras aplicar el test mencionado para la variable *age* obtenemos un p-valor inferior al test de significancia establecido por lo que podemos determinar que la variable *Age* no sigue una distribución normal.
- También aplicamos el test de Shapiro-Wilk para la variable *fare*. Se vuelve a obtener un p-valor inferior al nivel de significancia por lo que también podemos determinar que la variable *Fare* no sigue una distribución normal.

Dado que ambas variables no siguen una distribución normal, para estudiar la **homogeneidad** de la varianza se emplea el test de Fligner-Killen [6] suponiendo un nivel de significancia igual a 0.05. Se obtiene un p-valor inferior al nivel de significancia por lo que podemos determinar que las variables *Age* y *Fare* **presentan varianzas estadísticamente distintas.**

```
fligner.test(Age ~ Fare, data=df)

##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Fare
## Fligner-Killeen:med chi-squared = 319.57, df = 246, p-value = 0.001094
```

Ilustración 17 Resultado del test Fligner-Kilen para las variables age y fare

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

La supervivencia es mayor en mujeres que en hombres

A partir de un diagrama de barras podemos observar que la mayor parte de las personas que sobrevivieron al naufragio del Titanic eran mujeres. Comprobaremos este hecho con el diseño de un modelo de regresión logística con variable dependiente Survived y variable explicativa Sex.

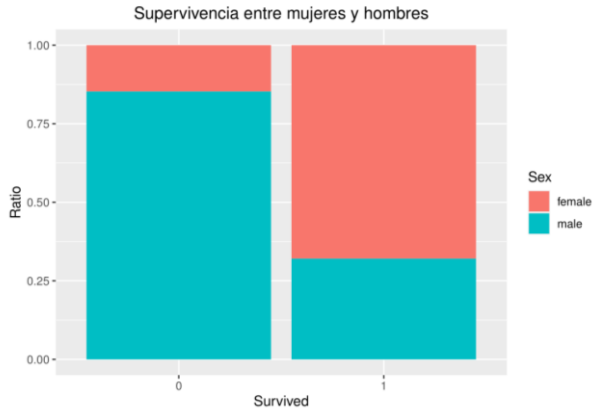

Ilustración 18 Comparativa de la supervivencia entre hombres y mujeres

Creamos el modelo de regresión logística y obtenemos un coeficiente asociado al valor male de la variable Sex negativo (columna Estimate) lo que indica que la probabilidad de sobrevivir al naufragio del Titanic era menor en hombres que en mujeres.

```
model_surv_sex <- glm(as.factor(Survived)~Sex, family=binomial(link=logit), data=df)
summary(model_surv_sex)
```

```
##
## Call:
## glm(formula = as.factor(Survived) ~ Sex, family = binomial(link = logit),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6423  -0.6471  -0.6471   0.7753   1.8256
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0480     0.1291   8.116 4.83e-16 ***
## Sexmale      -2.5051     0.1673 -14.975  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.82  on 888  degrees of freedom
## Residual deviance:  916.61  on 887  degrees of freedom
```

Ilustración 19 Resultados del modelo logístico entre las variables survived y sex

La supervivencia es mayor en los niños (La proporción de niños que sobrevivieron es mayor que la proporción de adultos que sobrevivieron)

Para determinar si la supervivencia es mayor en viajeros en los niños se empleará un contraste de hipótesis sobre la proporción de dos muestras (niños y adultos). Se consideran niños a toda aquella persona menor de 18 años.

Se definen las hipótesis nula y alternativa siguientes:

- Hipótesis nula:La proporción de niños que sobrevivieron es igual a la proporción de adultos.
  ρ Supervivientes(kids) = ρ Supervivientes(adults)
- Hipótesis alternativa:La proporción de niños que sobrevivieron es mayor a la proporción de adultos.
  ρ Supervivientes(kids) >ρ Supervivientes(adults)

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(kids.surv, adults.surv) out of c(n.kids, n.adults)
## X-squared = 9.3391, df = 1, p-value = 0.001122
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.06375797 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.5078125 0.3613666
```

Ilustración 20 Contraste de hipótesis de la proporción de supervivientes entre adultos y niños

Realizamos el contraste de hipótesis con la función **prop.test** y obtenemos un p-value igual a 0.001122. Se emplea un porcentaje de confianza del 95 por ciento por lo que el nivel de significancia es igual a 0.05 (1 - 95/100). Debido a que el p-value es inferior que el nivel de significancia podemos rechazar la hipótesis nula, aceptar la hipótesis alternativa y afirmar con un 95 por ciento de confianza que la proporción de los niños que sobrevivieron es mayor que la proporción de adultos.

La supervivencia es mayor en viajeros de primera clase

En primer lugar observamos gráficamente la proporción de pasajeros de cada clase que sobrevivieron y los que no. La proporción de pasajeros de la primera clase que no sobrevivieron es menor que la proporción del resto de clases. Por otro lado, la proporción de personas de la primera clase que sobrevivieron es la mayor de las tres clases.

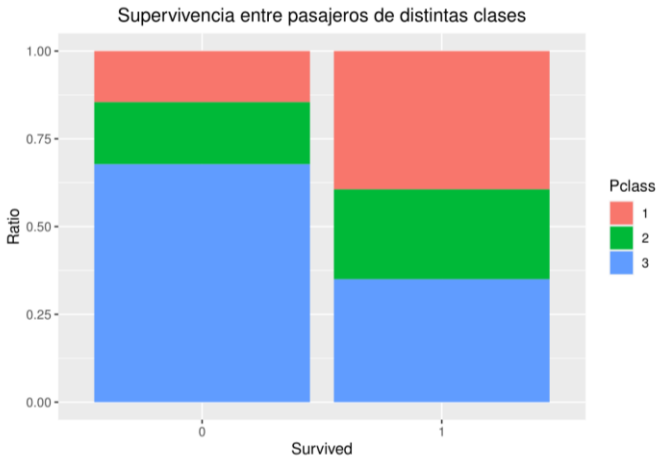

Ilustración 21 Comparativa de la variable supervivencia entre pasajeros de distintas clases

Para determinar si la supervivencia es mayor en viajeros de primera clase generamos un modelo de regresión logística y calculamos los **odds-ratios.** Se toma el valor 1 (primera clase) como valor de referencia para la variable explicativa. Los odds-ratios nos indican la diferencia entre la probabilidad de sobrevivir al naufragio si el pasajero viaja en primera clase (valor de referencia) y la probabilidad de sobrevivir si el pasajero viaja en la segunda y en la tercera clase.

```
Call:
glm(formula = as.factor(Survived) ~ Pclass, family = binomial(link = logit),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4028  -0.7450  -0.7450   0.9676   1.6836

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5158     0.1413   3.651 0.000261 ***
Pclass2      -0.6246     0.2044  -3.056 0.002241 **
Pclass3      -1.6556     0.1762  -9.395  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.8  on 888  degrees of freedom
Residual deviance: 1081.2  on 886  degrees of freedom
AIC: 1087.2

Number of Fisher Scoring iterations: 4
```

Ilustración 22 Resultados del modelo logístico entre las variables survived y pclass

Los odds-ratios se pueden obtener a partir de los coeficientes del modelo de regresión. Obtenemos dos odds-ratio distintos uno para la segunda clase y otro para la tercera clase.

```
exp(coefficients(model_surv_class))

## (Intercept)     Pclass2     Pclass3
##   1.6750000   0.5354670   0.1909806
```

Ilustración 23 Odds-ratio entre la variable suvived y las clases segunda y tercera

- El odds-ratio de la **segunda clase** es igual a 0.53. El hecho de obtener un odd-ratio menor que uno indica que la probabilidad de sobrevivir viajando en segunda clase es menor que si se viaja en primera clase.
- El odds-ratio de la **tercera clase** también es inferior a la unidad por lo que podemos determinar que la probabilidad de sobrevivir viajando en la tercera clase es menor que si se viaja en la primera clase.

Por todo ello podemos concluir que **la supervivencia era mayor si se viaja en la primera clase.**

Existe relación de supervivencia en niños y clase

Se quiere evaluar si existe correlación en el factor de ser niño y viajar en primera, segunda o tercera clase.

```
> cor(df.corr)
             Survived     Pclass         Age
Survived  1.00000000 -0.3355489 -0.08404038
Pclass   -0.33554886  1.0000000 -0.41025391
Age      -0.08404038 -0.4102539  1.00000000
```

Ilustración 24 Correlación entre las variables survived pclass y age

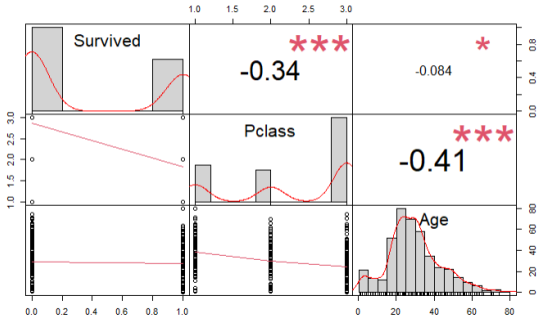

Ilustración 25 Relaciones entre las clases survived, pclass y age

En general se observa que parece no haber buenas relaciones lineales entre las variables, no hay evidencia de relaciones cercanas a -1 o 1, aunque tenemos algunos coeficientes de relación que no son buenos pero existe relación lineal: *PClass* con *Age* (-0.41), *Survived* con *PClass*(-0.34) Si observamos la gráfica generada con R, nos muestra en un tamaño de fuente superior las correlaciones con mejor coeficiente.

Analizaremos si existen relaciones utilizando la regresión logística:
Agrupamos a los pasajeros por edad: adulto (+18 años), adolescente(11-17 años) y niños (0 -10 años).

```
Call:
glm(formula = as.factor(Survived) ~ Pclass + age_label, family = binomi
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9320  -0.9277  -0.6587   0.9896   1.8081

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.6982     0.2968   5.721 1.06e-08 ***
Pclass2            -0.7286     0.2077  -3.508 0.000451 ***
Pclass3            -1.8769     0.1854 -10.122  < 2e-16 ***
age_labelteenager  -0.4416     0.3350  -1.318 0.187415
age_labeladult     -1.2389     0.2642  -4.689 2.75e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.8  on 888  degrees of freedom
Residual deviance: 1052.6  on 884  degrees of freedom
AIC: 1062.6

Number of Fisher Scoring iterations: 4
```

Ilustración 26 Resultados del modelo logístico entre las variables survived y las variables pclass y age

Se observa que las variables son Pclass y Age son significativas porque el p-value de cada una de ellas es <0,05 proporcionado por el estadístico de Wald. El p-value de cada variable predictora rechaza la hipótesis nula (coeficiente = 0). Podemos asumir que las variables predictoras tienen influencia en el modelo predictivo ya que cambiarán los valores en la variable dependiente *Survived*.

```
```{r}
exp(coefficients(model_surv_class))
```

     (Intercept)   Pclass2   Pclass3 age_labelteenager   age_labeladult
       5.4641629   0.4825677   0.1530621   0.6429938   0.2897024
```

Ilustración 27 Odds-ratio entre la variable survived, las clases segunda y tercera y distintos tramos de edad

En resumen, podemos asumir que si las probabilidades de supervivencia de ser niño y viajar en tercera clase son menores que si viaja en las otras clases.

```
> #Predicción ser niño (<10 años) y viajar en primera clase
> prediction_data <- list("age_label" = "kid", "Pclass" = "1")
> predict(model_surv_class, newdata = prediction_data, type="response")
        1
0.8453009
>
> #Predicción ser niño (<10 años) y viajar en segunda clase
> prediction_data <- list("age_label" = "kid", "Pclass" = "2")
> predict(model_surv_class, newdata = prediction_data, type="response")
        1
0.7250352
>
> #Predicción ser niño (<10 años) y viajar en tercera clase
> prediction_data <- list("age_label" = "kid", "Pclass" = "3")
> predict(model_surv_class, newdata = prediction_data, type="response")
        1
0.4554434
>
```

Ilustración 28 Cálculo de las probabilidades de sobrevivir si un niño pertenece a cada una de las clases

Interpretación:

- 84,5% probabilidades de sobrevivir si es niño de primer clase
- 72,5% probabilidades de sobrevivir si es niño de segunda clase
- 45,5 % probabilidades de sobrevivir si es niño de tercera clase

Las variables has_cabine y Pclass están relacionadas o son independientes

Para abordar si existen diferencias en la proporción de Pclass según si tiene cabina o no es realizando un test de independencia de dos variables cualitativas. Concretamente, nos preguntamos si el tener cabina y la clase están relacionadas o se pueden considerar variables independientes.

Hipótesis nula es H0: has_cabina y Pclass son variables independientes
Hipótesis alternativa es H1: existe una relación entre las variables has_cabina y Pclass

```
        Pearson's Chi-squared test

data:  df.chi
X-squared = 3932.8, df = 3552, p-value = 6.114e-06
```

Ilustración 29 Resultado del test chi-cuadrado entre las variables has_cabined y pclass

El test obtiene un p-valor 0,000006114 y es menor que el nivel de significancia establecido (0.05). Por lo tanto, si las frecuencias observadas se alejan significativamente de los valores esperados, podremos concluir que no existe una relación de dependencia entre las variables. Podemos rechazar la hipótesis nula por lo que hay suficiente evidencia para afirmar que no existe dependencia entre tener cabina y la clase en la que viaja con un nivel de confianza del 95%.

Existe relación entre el precio del billete y el puerto de embarque

Volvemos a revisar la distribución de los precios de los billetes y vemos que la media de precio fue 32.097 libras esterlinas y que hubo algún pasajero que llegó a pagar 512.329 libras esterlinas. Como tenemos valores extremos que son precios de billetes excepcionales, nos estropean la media, por lo que hemos de trabajar nuestros rangos sin tener en cuenta estos *outliers*.

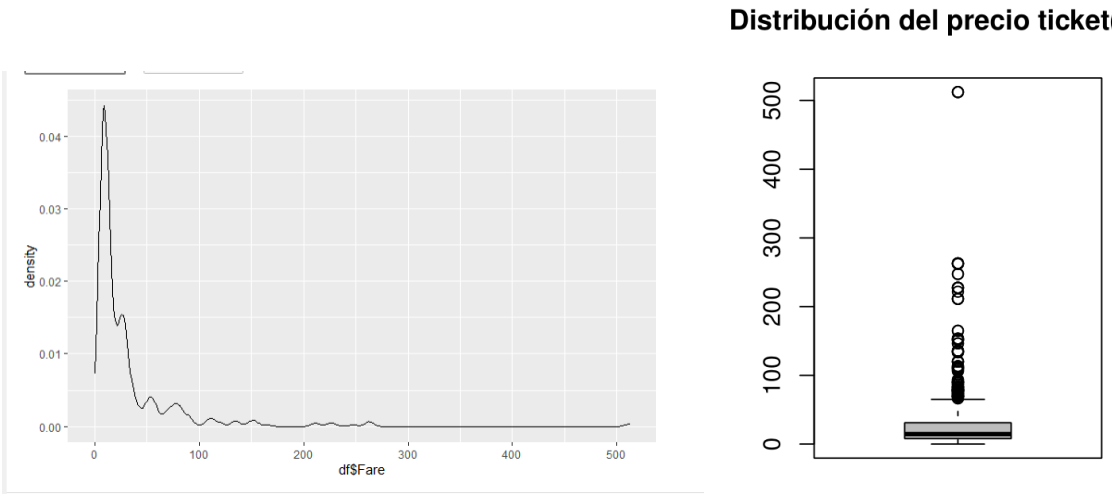

Ilustración 30 Distribución de la variable fare

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   7.896  14.454  32.097  31.000 512.329
```

Ilustración 31 Parámetros estadísticos de la variable fare

Rangos de precio:

- Económico: Entre 0 y 8.000
- Normal: 8.000 y 70.000
- Excesivo: Más de 70.000

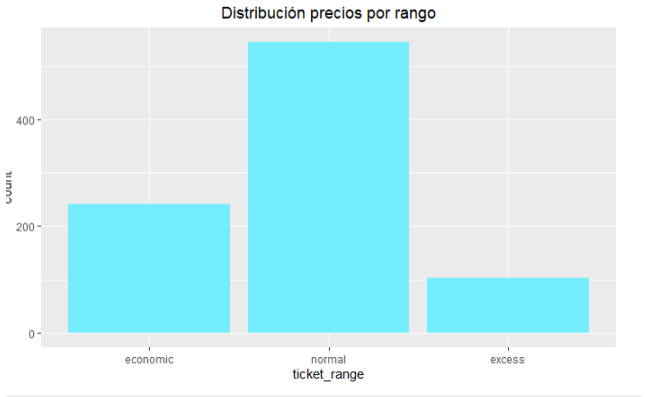

Ilustración 32 Distribución de precios por rango

Para abordar si existen diferencias en los precios del billete según el puerto de embarque realizamos un test de independencia de dos variables cualitativas.

```
        Pearson's Chi-squared test

data:  df.chi
X-squared = 248.16, df = 888, p-value = 1
```

Ilustración 33 Test de chi-cuadrado entre el precio del billete y el puerto de embarque

El test obtiene un p-valor 1 y es mayor que el nivel de significancia establecido (0.05). Por lo tanto, si las frecuencias observadas se alejan significativamente de los valores esperados, podremos concluir que existe una relación de dependencia entre las variables. Podemos

aceptar la hipótesis nula por lo que hay suficiente evidencia para afirmar que existe dependencia entre el precio del ticket y el puerto de embarque con un nivel de confianza del 95%.

La supervivencia es mayor en pasajeros con cabina

Agrupamos a los pasajeros según el atributo has_cabine que habíamos creado al principio del análisis del conjunto de datos, los pasajeros “Con cabina” y “Sin cabina” y generamos el modelo logístico para ver si existe correlación entre los atributos.

```
Call:
glm(formula = as.factor(Survived) ~ has_cabin, family = binomial(link = logit),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4756  -0.8444  -0.8444   0.9060   1.5521

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.6783     0.1489   4.556 5.22e-06 ***
has_cabinSin cabina -1.5263     0.1706  -8.947  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.8  on 888  degrees of freedom
Residual deviance: 1097.2  on 887  degrees of freedom
AIC: 1101.2

Number of Fisher Scoring iterations: 4

        1
0.6633663
        1
0.2998544
```

Ilustración 34 Resultados del modelo logístico entre las variables survived y has_cabine

Si utilizamos el modelo para realizar dos predicciones de supervivencia de un pasajero con cabina obtenemos 0,6633 y para la predicción de un pasajero sin cabina se obtiene 0,2998 que son las probabilidades de la variable dependiente, por lo que podemos interpretar que existe un:

- 66,3% probabilidades de sobrevivir si el pasajero tenía cabina asignada
- 29,9% probabilidades de sobrevivir si el pasajero no tenía cabina asignada

Por lo que creemos que quizá los pasajeros con cabinas tenían mejor accesibilidad a los botes salvavidas que los que no tenían cabina.

La supervivencia menor en pasajeros que viajan solos

La última relación que nos quedará analizar es si la probabilidad de sobrevivir es menor si se viajaba solo en el Titanic, es decir sin hermanos/cónyuges (SibSp) o hijos/padres (Parch) a bordo.

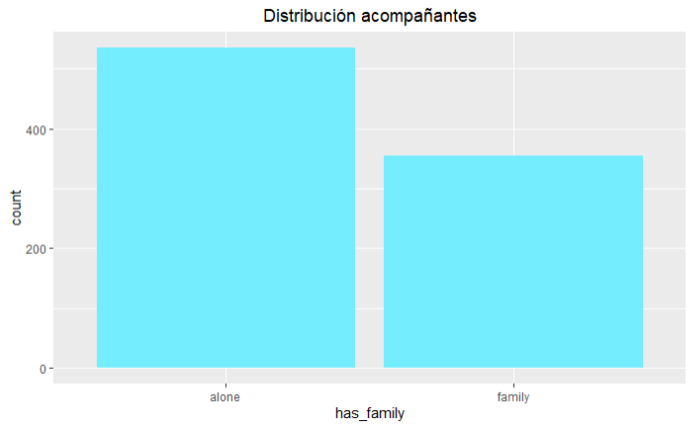

Ilustración 35 Comparativa entre las personas que viajaban solas o con familia

```
          1
0.3009346
          1
0.5056497
```

Ilustración 36 Resultado del cálculo de la probabilidad de sobrevivir si se viajaba solo o en familia.

Obtenemos que la probabilidad de sobrevivir si se viajaba solo es de 30% frente a 50% si se viajaba con algún familiar acompañante. Esto tiene sentido, ya que el acompañante podría ser un niño, y dejaban acceder a los botes salvavidas a un adulto por niño.

5. Conclusiones.

Una de las primeras conclusiones es que **sobrevivieron tres veces más mujeres que hombres**. Una explicación podría deberse al protocolo de evacuación “Mujeres y niños primero”, y así fue en el Titanic, pero no por protocolo, ya que entre el pánico del naufragio fue el de “sálvese quien pueda”, si no por la orden del capitán de disparar a todo el que no diera paso a mujeres y niños en la entrada a los botes salvavidas.

Por otro lado, la edad de los pasajeros es bastante similar entre supervivientes y fallecidos, con la excepción de **los niños menores de 10 años** donde el porcentaje de supervivencia es superior al de toda la muestra.

Por otra parte, los pasajeros con hijos/padres a bordo fueron las más afortunadas. Con ninguna familia, solo un tercio sobrevivió. En el caso de familiares no sanguíneos o esposa/o, se cumple una relación parecida: **las personas con un familiar** sobrevivieron más que las que iban solas.

Cruzando las variables *Survived*, *Age* y *Pclass* observamos que **los niños de tercera clase representan gran parte de los menores que perdieron la vida en la catástrofe**. En primera clase solo había cuatro niños, de los cuales murió uno. Los niños de segunda clase (17) , sobrevivieron. Las personas más **ancianas murieron**. En los **adultos** (entre 20 y 50 años), las personas de condición más **humilde representaron la mayor proporción de los muertos**.

Los billetes de primera clase costaban mucho más dinero que los de la segunda y tercera clase, por lo que podemos inferir que **el precio medio de los tickets** de las personas que **sobrevivieron era superior al de los que fallecieron**.

Gracias al procesamiento de este conjunto de datos, podemos construir modelos de predicción con algoritmos de *machine learning* que permitirán decirnos con precisión la probabilidad de supervivencia dependiendo de los factores: sexo, clase y si tiene al menos un hijo a bordo.

Código.

El código R con el que se realizado la limpieza del dataset y el estudio estadístico, se encuentra disponible en el repositorio Git: https://github.com/Angeles1/Titanic-CleanData/

Referencias

[1] Wikipedia, «RMS Titanic,» [En línea]. Available: https://es.wikipedia.org/wiki/RMS_Titanic.

[2] Kaggle, «Kaggle,» [En línea]. Available: https://www.kaggle.com/c/titanic.

[3] National Geographic, «National Geographic,» [En línea]. Available: https://historia.nationalgeographic.com.es/a/vidas-truncadas-titanic_11387.

[4] SOY505, «Este fue el precio de los boletos para viajar en el Titanic,» [En línea]. Available: https://www.soy502.com/articulo/este-fue-precio-boletos-viajar-titanic-32539.

[5] Wikipedia, «Test de Shapiro–Wilk,» [En línea]. Available: https://es.wikipedia.org/wiki/Test_de_Shapiro%E2%80%93Wilk.

[6] GeeksforGeeks, «Fligner-Killeen Test in R Programming,» [En línea]. Available: https://www.geeksforgeeks.org/fligner-killeen-test-in-r-programming/.

Contribuciones	**Firma**
Investigación previa	MAFE NJCF
Redacción de las respuestas	MAFE NJCF
Desarrollo del código	MAFE NJCF