Alumnos:
Albert Breva
Juan Matilla

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

El hundimiento del titanic es seguramente el naufragio más famoso de la historia. El 15 de abril de 1912, durante su viaje inaugural, el Titanic se hundió después de chocar contra un iceberg. Desafortunadamente, no había botes salvadidas para todo el mundo por lo que fallecieron muchos de sus pasajeros y la tripulación.
La pregunta que se pretende resolver en este estudio es cuales fueron las causas que permitieron a unos pasajeros sobrevivir atendiendo a variables de género, socioeonómicas, etcétera.
El dataset *titanic* que vamos a utilizar se ha extraído de kaggle y contiene datos estructurados en 891 filas y 12 columnas y con el esperamos poder responder a la cuestión previamente planteada.

En la siguiente tabla podemos ver las principales características de las distintas variables de nuestro conjunto de datos:

> *summary(titanic)*

```
  PassengerId       Survived          Pclass                                      Name         Sex          Age            SibSp
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                  :  1   female:314   Min.   : 0.42   Min.   :0.000
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward          :  1   male  :577   1st Qu.:20.12   1st Qu.:0.000
 Median :446.0   Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)     :  1                Median :28.00   Median :0.000
 Mean   :446.0   Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                  :  1                Mean   :29.70   Mean   :0.523
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00   3rd Qu.:1.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin       :  1                Max.   :80.00   Max.   :8.000
                                                  (Other)                              :885                NA's   :177
     Parch             Ticket         Fare               Cabin     Embarked
 Min.   :0.0000   1601     :  7   Min.   :  0.00             :687    :  2
 1st Qu.:0.0000   347082   :  7   1st Qu.:  7.91   B96 B98   :  4   C:168
 Median :0.0000   CA. 2343:  7   Median : 14.45   C23 C25 C27:  4   Q: 77
 Mean   :0.3816   3101295  :  6   Mean   : 32.20   G6        :  4   S:644
 3rd Qu.:0.0000   347088   :  6   3rd Qu.: 31.00   C22 C26   :  3
 Max.   :6.0000   CA 2144  :  6   Max.   :512.33   D         :  3
                  (Other)  :852                    (Other)   :186
```

A continuación se mostrará un breve resumen de que significan las distintas variables:
PassangerId: Numeración de los pasajeros; del 1 al 891.
Survived: 0 si el pasajero no sobrevivió y 1 si sí lo hizo.
Pclass: Clase del pasajero; 1 para 1ª clase, 2 para 2ª y 3 para los de 3ª clase.
Name: Nombre de los pasajeros.
Sex: Género de los pasajeros.
Age: Edad de los pasajeros.
SibSp: Número de esposas y/o hermanos a bordo.
Parch: Número de padres y/o hijos a bordo.
Ticket: Identificativo del billete de los pasajeros.
Fare: Precio de los tickets de los pasajeros.
Cabin: Cabina en la que viajaban el pasajeros.
Embarked: Lugar en el que embarcaron los pasajeros.

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

Teniendo en cuenta el análisis previo de los datos mostrado en el apartado anterior, hay unas variables que a primera vista podemos concluir que no tendrán relación con la supervivencia del pasajero.
Estás variables son únicas para cada pasajero y, por tanto, no permiten hacer agrupaciones de diversos pasajeros ni determinar causas que favoreciesen la supervivencia.
Las variables que se eliminan son: PassangerId, Name y Ticket.

3. Limpieza de datos.

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

Nuestros datos contienen elementos vacíos que no podemos recuperar puesto que no se trata de una base de datos nuestra. Muchas veces los datos perdidos vienen directamente como NA, pero otras veces pueden ser un elemento vacío, un carácter extraño, etcétera, por lo que creamos una función que nos permita catalogar como NA's muchos de estos casos y así asegurarnos de no dejar valores perdidos en nuestro dataset.

```
> unify_null <- function(x){
 na_index<-(is.na(x) | x=="null" | x=="NULL" | x=='\\N' | x=='\\n' | x=="" | x=="?" |
x=='NA');
 x[na_index]<-NA;
 return(x)
}
```

Una vez aplicada la función y calculado un porcentaje podemos ver cuales de nuestras variables tienen un porcentaje mayor de valores perdidos, siendo Age, Cabin y Embarked las únicas variables que tienen este problema, aunque en distinto grado.

```
> titanic <- data.table(apply(titanic, 2, unify_null))
> sapply(titanic, function(x){100*sum(is.na(x))/length(x)})
```

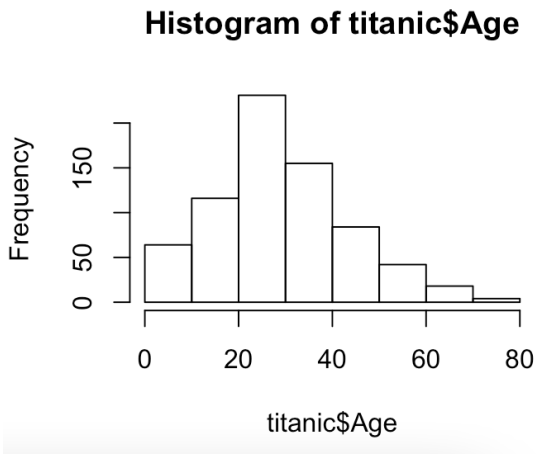
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	19.8653199	0.0000000	0.0000000	0.0000000	0.0000000	77.1043771	0.2244669

Vemos como la variable Cabin tiene un 77% de valores perdidos por lo que decidimos eliminarla y no realizar en este caso ningún proceso de imputación de valores faltantes.
En el caso de la variable Age, que casi un 20% de valores perdidos, lo primero que hacemos es visualizar sus estadísticos y realizar un sencillo histograma.

```
> titanic$Age <- as.integer(titanic$Age)
> summary(titanic$Age, na.rm = TRUE)
```

```
> hist(titanic$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   20.00   28.00   29.68   38.00   80.00     177
```



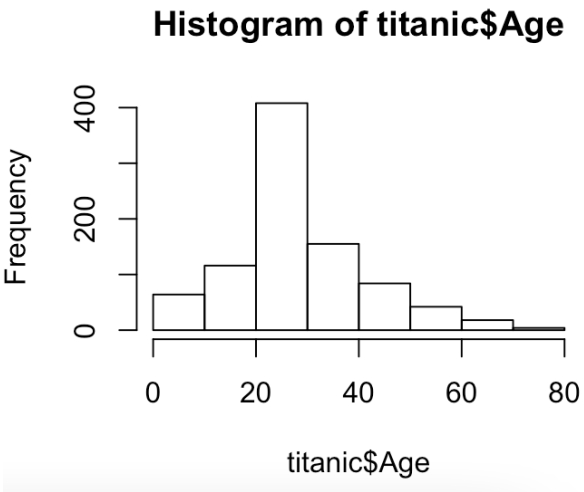
Observando estos valores decidimos reemplazar los valores perdidos por la media.

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

Si visualizamos de nuevo los estadísticos y el histograma vemos que los datos han cambiado ligeramente aunque la distribución no se ha visto demasiado alterada.

```
> summary(titanic$Age, na.rm = TRUE)
> hist(titanic$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   22.00   29.68   29.68   35.00   80.00
```



Por ultimo, veamos el caso de los valores perdidos en la variable Embarked, donde apenas representaban un 0,22%.

```
> titanic$Embarked <- as.factor(titanic$Embarked)
> summary(titanic$Embarked, na.rm = TRUE)
```

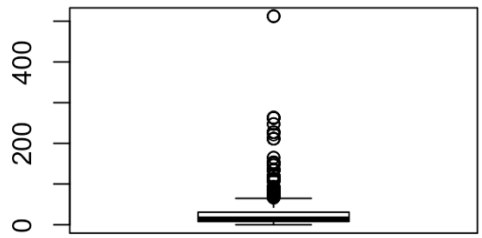
```
  C   Q   S NA's
168  77 644   2
```

Como vemos solamente temenos 2 NA's en esta variable así que los imputamos a la categoría que más aparece, en este caso la 'S'.

```
> titanic$Embarked[is.na(titanic$Embarked)] <- 'S'
```

3.2. Identificación y tratamiento de valores extremos.

Para tratar nuestros valores extremos lo primero que hacemos son una serie de transformaciones muy sencillas del tipo de nuestras variables y a continuación realizamos unas sencillas visualizaciones que nos permitan detectar valores outliers. Viendo el boxplot resultante de la variable Fare parece evidente que tenemos un problema de outliers.



Lo que hacemos es visualizar los estadísticos, almacenar el boxplot en un objeto y a continuación utilizar el parámetro out de R que nos dice cuales son los valores que se consideran outliers.

```
> summary(titanic$Fare)
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    7.91   14.45   32.20   31.00  512.33
```

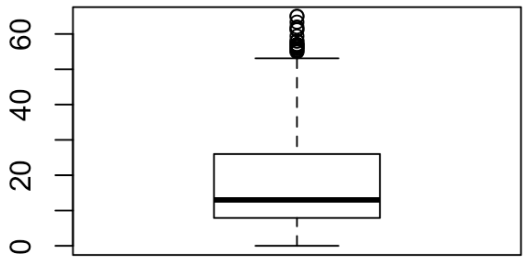
Aunque ya se veía en el boxplot, viendo estos estadísticos con un valor máximo de 512,33 es claramente obvio que tenemos un problema de outliers.

```
> fare_plot <- boxplot(titanic$Fare)
> fare_plot$out
```

```
 [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000
[15]  66.6000  69.5500  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000  79.2000  86.5000 512.3292  79.6500
[29] 153.4625 135.6333  77.9583  78.8500  91.0792 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750 164.8667 134.5000
[43]  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000 263.0000  75.2500  69.3000 135.6333  82.1708 211.5000 227.5250  73.5000
[57] 120.0000 113.2750  90.0000 120.0000 263.0000  81.8583  89.1042  91.0792  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000
[71] 221.7792 106.4250  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000  78.2667 153.4625  77.9583  69.3000
[85]  76.7292  73.5000 113.2750 133.6500  73.5000 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250 211.3375 512.3292
[99]  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583 211.3375  79.2000  69.5500 120.0000  93.5000  80.0000  83.1583  69.5500
[113]  89.1042 164.8667  69.5500  83.1583
```

Como podemos ver, tenemos más de 100 valores outliers en esta variable así que es conveniente eliminarlos, aunque se reduzca el número de observaciones de nuestro dataset, porque si no seguramente afectarán a las comprobaciones y modelos futuros.
De tener un dataset lo suficientemente grande podríamos plantearnos dejar los valores extremos, pero dada la limitación de observaciones que tenemos en el conjunto de datos original y el elevado número de outliers en esta variable es claro que nos afectarían y por eso decidimos eliminarlos.

```
> df<-df[!(df$Wind %in% g_caja$out),]
> titanic <- titanic[!(titanic$Fare %in% fare_plot$out),]
> boxplot(titanic$Fare)
```




```
> summary(titanic$Fare)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   7.896  13.000  17.822  26.000  65.000
```

Como podíamos ver una vez eliminados los valores extremos ahora la distribución en el boxplot es mucho mejor y los estadísticos son mucho más razonables.

4. Análisis de los datos.

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

Analizaremos los datos para los pasajeros de clase baja para ver en los posteriores análisis si estos pasajeros tenían mas o menos posibilidades de sobrevivir al accidente del titanic.
Para ello, creamos una variable dicotómica TRUE o FALSE sobre si un pasajero es de clase baja o no.

```
> titanic$Clase_baja <- ifelse(titanic$Pclass == 3, TRUE, FALSE)
```

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

Para comprobar la normalidad de los datos podemos utilizar el test de Shapiro-Wilk o el test de Kolmogorv-Smirnov en las dos variables numéricas que tenemos en nuestro dataset, que son Age y Fare.

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

        Shapiro-Wilk normality test

data:  titanic$Age
W = 0.9541, p-value = 0.000000000000008048
```

Según el test de Shapiro-Wilk tenemos un p-value menor al nivel de significancia α = 0,05 por lo que podemos rechazar la hipótesis nula y concluir que los datos de la variable Age no cuentan con una distribución normal.

```
> ks.test(titanic$Age, pnorm, mean(titanic$Age), sd(titanic$Age))

        One-sample Kolmogorov-Smirnov test

data:  titanic$Age
D = 0.15398, p-value = 0.00000000000000222
alternative hypothesis: two-sided
```

El test de Kolmogorov-Smirnov nos da también un p-value menor que 0,05 por lo que podemos concluir definitivamente que la variable no está distribuida normalmente.

Veamos ahora la distribución de la variable Fare con estos mismos test.

```
> shapiro.test(titanic$Fare)

        Shapiro-Wilk normality test

data:  titanic$Fare
W = 0.81301, p-value < 0.00000000000000022
```

```
> ks.test(titanic$Fare, pnorm, mean(titanic$Fare), sd(titanic$Fare))

        One-sample Kolmogorov-Smirnov test

data:  titanic$Fare
D = 0.19101, p-value < 0.00000000000000022
alternative hypothesis: two-sided
```

Como Podemos observer ambos test nos dan un p-value menor que 0,05 por lo que podemos rechazar la hipótesis nula y concluir que los datos de la variable Fare no siguen una distribución normal.

A continuación vamos a comprobar la homocedasticidad en los datos, es decir, la igualdad de varianzas entre los grupos que se van a comparar. Como ya hemos comprobado previamente que los datos no siguen una distribución normal utilizaremos a continuación el test de Fligner-Killen.

```
> fligner.test(Age ~  Survived, data = titanic)

        Fligner-Killeen test of homogeneity of variances

data:  Age by Survived
Fligner-Killeen:med chi-squared = 3.3102, df = 1, p-value = 0.06885
```

```
> fligner.test(Fare ~ Survived, data = titanic)

        Fligner-Killeen test of homogeneity of variances

data:  Fare by Survived
Fligner-Killeen:med chi-squared = 35.945, df = 1, p-value = 0.000000002029
```

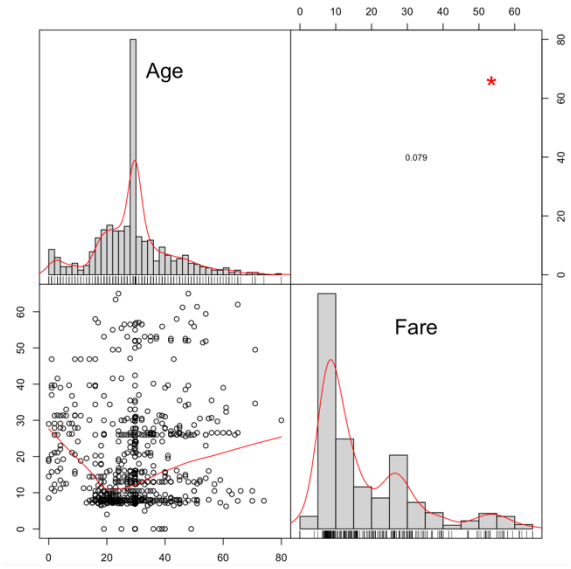
De los resultados de estos dos test podemos concluir que la varianza de Age es similar entre supervivientes y no supervivientes, pero diferente en el caso de la variable Fare.

4.3. 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.

Dado que los datos a analizar no presentan una distribución normal, las pruebas estadísticas para analizar y comparar los diferentes grupos deberán ser no parámetricas.

En este caso, la primera prueba que realizaremos es un análisis de correlaciones entre las variables Age y Fare.

```
> correlaciones <- titanic[, c(4,7)]
> chart.Correlation(correlaciones, histogram = TRUE, method = "pearson")
```



Tal y como podemos observar, no parece existir una correlación entre la edad de los pasajeros que iban a bordo del Titanic y la tarifa de su billete.

A continuación, realizaremos un contraste de hipótesis teniendo en cuenta el precio del billete de los que sobreviven y de los que no sobreviven al accidente.

```
> sobrevive <- titanic %>% filter(Survived == 1) %>% pull(Fare)
> nosobrevive <- titanic %>% filter(Survived == 0) %>% pull(Fare)
```

```
> wilcox.test(x=sobrevive,y=nosobrevive, paired = F)

        Wilcoxon rank sum test with continuity correction

data:  sobrevive and nosobrevive
W = 87641, p-value = 0.000000000005741
alternative hypothesis: true location shift is not equal to 0
```

Viendo los datos que nos devuelve el test de Wilcoxon podemos ver que las diferencias son estadísticamente significativas (p<0,05) por lo que podemos afirmar que había una diferencia significativa en el precio de los billetes de aquellos que sobrevivieron y de aquellos que fallecieron.

```
> mean(sobrevive)
[1] 22.26036
> mean(nosobrevive)
[1] 15.54228
> mean(sobrevive) – mean(nosobrevive)
[1] 6.718087
```

Como podemos ver, el precio medio de los billetes de los supervivientes estaba en torno a los 22$, el de aquellos que fallecieron en el accidente rondaba los 15$ y la diferencia era de casi 7$.

El último análisis estadístico que realizaremos es una regresión logística. Para ello creamos índices para dividir la base en entrenamiento y test con una misma relación de la variable objetivo entre ellas y una división de 80-20.

```
> train_index <- createDataPartition(y = titanic$Survived, p = 0.8, list = FALSE, times = 1)
> dat_train <- titanic[train_index, ]
> dat_test  <- titanic[-train_index, ]
> train_index <- sample(1:nrow(titanic), 0.8*nrow(titanic))
> dat_train <- titanic[train_index, ]
> dat_test  <- titanic[-train_index, ]

> train_label <- dat_train$Survived
> test_label <- dat_test$Survived
```

Lo siguiente es crear nuestro modelo de regresión en el cual intentamos predecir la supervivencia de los pasajeros del Titanic.

```
> modelo1<-glm(Survived ~ Clase_baja +  Age + Fare, data=dat_train, na.action = "na.omit", family = "binomial")
> summary(modelo1)

Call:
glm(formula = Survived ~ Clase_baja + Age + Fare, family = "binomial",
    data = dat_train, na.action = "na.omit")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6194  -0.7905  -0.6584   1.0812   2.3037

Coefficients:
               Estimate Std. Error z value       Pr(>|z|)
(Intercept)     0.838116   0.348288   2.406         0.0161 *
Clase_bajaTRUE -1.321910   0.213325  -6.197 0.000000000577 ***
Age            -0.035679   0.007694  -4.637 0.000003533017 ***
Fare            0.015764   0.007069   2.230         0.0258 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 795.15  on 619  degrees of freedom
Residual deviance: 715.16  on 616  degrees of freedom
AIC: 723.16

Number of Fisher Scoring iterations: 4
```

Interpretemos los resultados de las variables significativas.

- Clase_bajaTRUE: el signo es negativo así que podemos decir que la prabibilidad de sobrevivir al accidente es menor cuando se pertenece a la clase baja que cuando no.

- Age: el coeficiente de -0,03 nos está indicando que el logaritmo de odds ratio de sobrevivir al accidente disminuye 0,03 cuando aumenta una unidad la variable Age.
- Fare: el coeficiente de 0,01 nos indica que el logaritmo de sobrevivir al accidente aumenta 0,01 cuando aumenta una unidad la variable Fare.

A continuación hacemos la predicción de nuestro modelo.

```
> pred_logistic <- predict(object = modelo1, newdata =  dat_test, type = "response")
> pred_label <- as.factor(ifelse(pred_logistic < 0.6, 0, 1))
> dat_test$Survived <- as.factor(dat_test$Survived)
```

Finalmente hacemos la evaluación del mismo.

```
> Log_eval <- confusionMatrix(data = pred_label, reference = test_label, positive = "1")
> Log_eval

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 102  43
         1   1   9

               Accuracy : 0.7161
                 95% CI : (0.6382, 0.7856)
    No Information Rate : 0.6645
    P-Value [Acc > NIR] : 0.09983

                  Kappa : 0.2042

 Mcnemar's Test P-Value : 0.000000000637

            Sensitivity : 0.17308
            Specificity : 0.99029
         Pos Pred Value : 0.90000
         Neg Pred Value : 0.70345
             Prevalence : 0.33548
         Detection Rate : 0.05806
   Detection Prevalence : 0.06452
      Balanced Accuracy : 0.58168

       'Positive' Class : 1
```

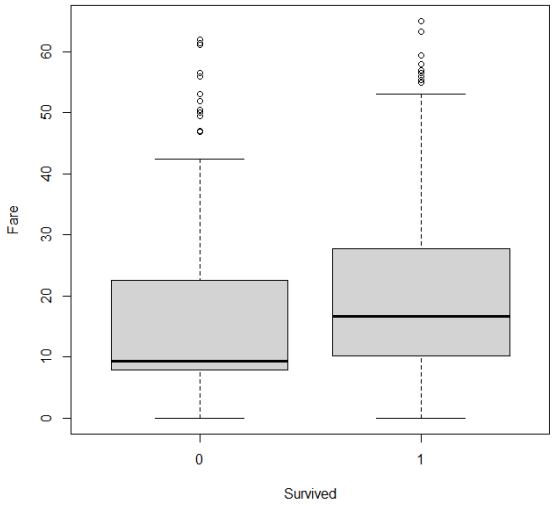
Veamos primero la matriz de confusión.
Lo primero que observamos es que el modelo clasifica correctamente 102 muestras de no supervivientes que efectivamente no lo son. Por otro lado, tenemos 43 falsos negativos lo que significa que nuestro modelo está diciendo que hay 43 pasajeros que fallecieron pero que en realidad sobrevivieron al accidente. Tenemos a su vez un falso

positivo, es decir un fallecido que el modelo predice que sobrevivió y 9 verdaderos negativos, es decir, 9 supervivientes correctamente clasificados.

En cuanto al accuracy es de 0,716 por lo que nuestro modelo predice con bastante exactitud que pasajeros sobrevivieron o fallecieron al accidente del titanic.
Por otro lado, la sensibilidad es de tan solo 0,17, mientras que la especificidad es de 0,99.

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

Como se comentó en el apartado 4.3, podemos observar mediante boxplots las diferentes distribuciones del precio del billete en función de si el pasajero sobrevivió o no.



Se puede observar cómo en función de la clase en la que viajaban los pasajeros fue relevante a la hora de que sobreviviesen al accidente. Cómo podemos observar en el siguiente gráfico, la proporción de supervivientes de clase 1 es mucho mayor que la proporción de clase 1 de los no supervivientes, al contrario que los de clase más baja (3), donde la proporción es mucho mayor que en los no supervivientes

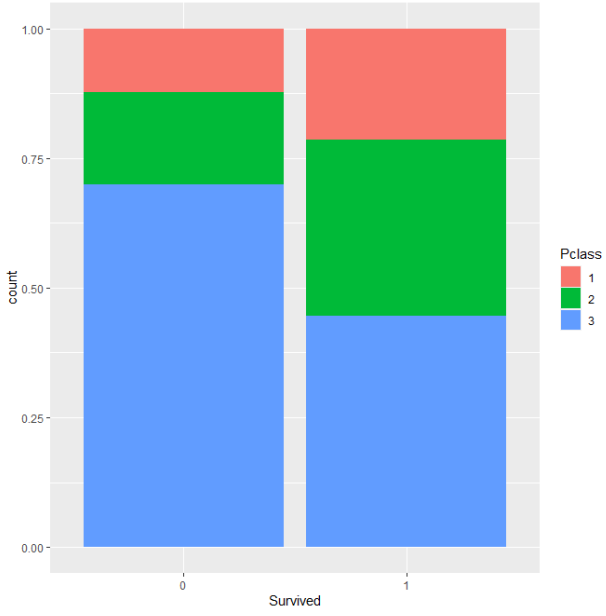


Tabla de resultados de la matriz de confusión utilizando la regresión logística.

Accuracy	0.72
Specifity	0.99
Precision	0.90
Recall	0.17

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

Los resultados obtenidos al utilizar el dataset de titanic nos permiten afirmar que efectivamente el precio del billete o la clase a la que pertenecían los viajeros fueron condicionantes para sobrevivir o no al accidente del titanic.
Nuestro modelo presenta una exactitud del 72% por lo que es bastante útil a la hora de predecir si una persona sobrevive o no a dicho accidente. Una de las limitaciones principales a la hora de trabajar con este conjunto de datos es la poca cantidad de registros, que además se ve disminuida al tener que eliminar un importante número de outliers. De contar con un dataset con más registros es muy probable que se pudiese mejorar el modelo y predecir con mayor exactitud que pasajeros hubiesen sobrevivido o fallecido en el accidente del Titanic.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Tabla de contribuciones

Contribuciones	**Firma**
Investigación previa	AB, JM
Redacción de las respuestas	AB, JM
Desarrollo del código	AB, JM