Tipología y ciclo de vida de los datos: Limpieza y análisis de datos

Autor: ROSEMBERG ALVAREZ DÍAZ

Mayo 2020

#Descripción del dataset

El dataset contiene registros sobre los pasajeros que abordaron el Titanic el 10 de abril de 1912 desde el puerto Southampton y los pasajeros que se incorporaron en los puertos de Cherburgo, Francia, y en Queenstown en Irlanda. Entre los pasajeros que abordaron se encontraban personas adineradas e inmigrantes que buscaban mejores opciones de vida al llegar Norteamérica. En el naufragio ocurrido en la madrugada del 14 de abril de 1912, fallecieron 1514 personas de las 2223 que abordaron. lo que convierte a esta tragedia en uno de los mayores naufragios de la historia. Es importante aclarar que los datos se encuentran divididos en dos (2) grupos: conjunto de entrenamiento (train.csv) y conjunto de prueba (test.csv), para esta practica integraremos ambos conjuntos. El conjunto de datos resultante está constituido por las siguientes variables:

- survived: Si la persona sobrevivió o no al naufragio 0 = No, 1 = Si.
- pcclass: Clase en la que viajaba la persona. 1 = primera, 2 = segunda, 3 = tercera.
- name: Nombre del pasajero.
- sex: Sexo del pasajero, male o famale.
- age: Edad del pasajero.
- sibsp: Número de hermanos que el pasajero tenía a bordo.
- parch: Número de padres (del pasajero) que estaban a bordo.
- ticket: Número de tikcet que el pasajero entregó al abordar.
- fare: Monto que el pasajero pago para obtener su boleto.
- cabin: Cabina que fue asignada al pasajero.
- embarked: Indica el puerto donde el pasajero abordo, C = cherbourg, Q = Queenstown, S= Southampton.

A partir de este conjunto de datos se plantea la problemática de determinar qué variables influyeron más sobre si un pasajero sobrevivió o no al naufragio. Además, se podrá proceder a crear modelos de regresión que permitan predecir si un pasajero sobrevive o no al naufragio en función de sus características y contrastes de hipótesis.

#Integración y selección de los datos

```
# Cargamos los paquetes R que vamos a usar
library(ggplot2)
library(dplyr)
library(VIM)
```

Procedemos a realizar la lectura del fichero de entrenamiento los cuales se encuentra formato CSV.

```
#Almacenamos el conjunto de datos test y train un objeto data.frame.
test <- read.csv('titanic-test.csv',stringsAsFactors = FALSE)
train <- read.csv('titanic-train.csv', stringsAsFactors = FALSE)
```

```
#Fusionamos los conjuntos de datos test y train en un único dataset a través de la función bind_rows.
Titanic <- bind_rows(train,test)
filas=dim(train)[1]
```

```r
#Verificamos la estructura del nuevo conjunto de datos.
str(Titanic)
```

```
## 'data.frame':    1309 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" ...
```

Al llevar a cabo el análisis de las variables presentes en el conjunto de datos, las variables más significativas son: survived, pcclas, fare, sexo y age La variable survival es imprescindible para conocer si el pasajero sobrevivió al naufragio, las variables pcclas y fare, pueden darnos noción de que tanto influye la posición económica al momento de rescatar a alguien, con las variables age y sex, podremos corroborar si se tuvo en cuenta la premisa en cualquier emergencia de dar prioridad a las mujeres y a los menores de edad.

```r
#Seleccionar las variables survived, pcclas, fare, sexo y age.
Titanic <- Titanic[,c(2,3,5,6,10)]

#Verificamos la estructura del nuevo conjunto de datos.
str(Titanic)
```

```
## 'data.frame':    1309 obs. of  5 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Sex     : chr  "male" "female" "female" "female" ...
##  $ Age     : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

#Limpieza de los datos

##Identificación y tratamiento de valores cero y vacíos.

```r
# Verificar la cantidad de valores null para cada uno de los atributos del juego de datos Titanic.
colSums(is.na(Titanic))
```

```
## Survived   Pclass      Sex      Age     Fare
##      418        0        0      263        1
```

Como podemos observar las variables Survived, Age y Fare contienen valores nulos que si no son tratados obtendremos análisis con alto margen de error. Para lo cual se empleará el método de la imputación basada en k vecinos más próximos.

```r
# Realizamos la imputación de valores mediante la función kNN() del paquete VIM.
Titanic$Survived <- kNN(Titanic)$Survived
Titanic$Age <- kNN(Titanic)$Age
Titanic$Fare <- kNN(Titanic)$Fare

# Verificar nuevamente si hay presencia de valores null para cada uno de los atributos del juego de dat
colSums(is.na(Titanic))
```

```
## Survived   Pclass      Sex      Age     Fare
##        0        0        0        0        0
```

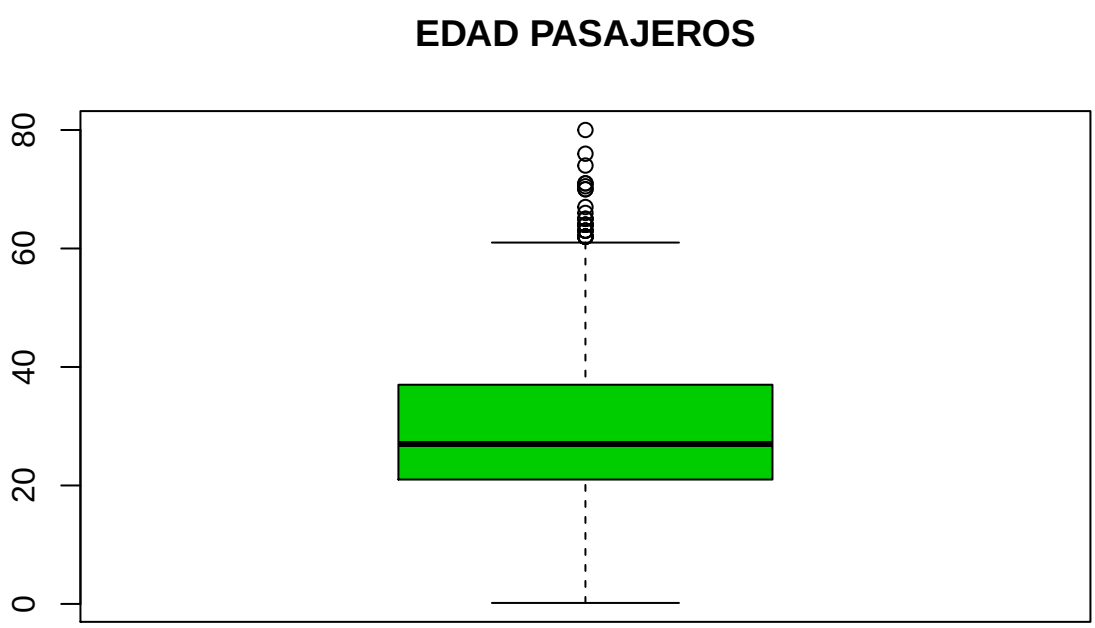
```
#Recodificar la variable Sexo (Sex). Si el pasajero es de sexo masculino se identificará con el numero
Titanic$Sex <- ifelse(Titanic$Sex == "male",1,0)

#Convertimos la variable Sexo (Sex) en un tipo entero.
Titanic$Sex <- as.integer(Titanic$Sex)
```

##Identificación y tratamiento de valores extremos

Se llevara a cabo la identificación de los valores extremos de las variables nuemericas Age y Fare a partir de la representacia de diagramas de caja (Boxplot) y la extración de los valores extremos a partir de la función boxplots.stats().

```
#Graficar valores extremos de la variable Age.
boxplot(Titanic$Age,
        col = c("51"),
        main = "EDAD PASAJEROS",
        yLab = "Años")
```



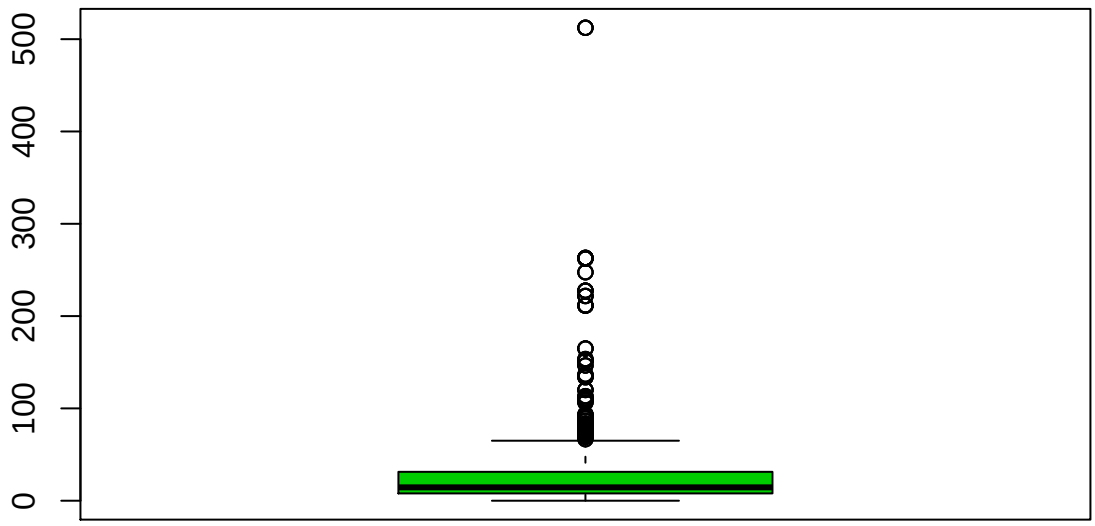

```
#Visualizamos los valores extremos de la variable edad (Age).
boxplot(Titanic$Age, col = c("51"), main="EDAD PASAJEROS", ylab="Años", horizontal = T, plot = F)
```

```
## $stats
##       [,1]
## [1,]  0.17
## [2,] 21.00
```

```
## [3,] 27.00
## [4,] 37.00
## [5,] 61.00
##
## $n
## [1] 1309
##
## $conf
##          [,1]
## [1,] 26.30127
## [2,] 27.69873
##
## $out
##  [1] 66.0 65.0 71.0 70.5 62.0 63.0 65.0 64.0 65.0 63.0 71.0 64.0 62.0 62.0 80.0
## [16] 70.0 62.0 70.0 62.0 74.0 62.0 63.0 62.0 67.0 76.0 63.0 62.0 64.0 64.0 64.0
##
## $group
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## $names
## [1] ""
```

Como podemos observar en el vector stats el ultimo valor que se obtuvo del límite inferior fue de 0.17 y de 61 para el límite superior y una mediana de 27. Existen treinta (30) outliers con unos valores que oscilan entre los 62 y 80 años. No es necesario llevar a cabo ningún tratamiento con los valores extremos de la variable Age, ya que la edad para algunos pasajeros podría oscilar entre los 64 y 80 Años al momento de abordar, lo cual es muy viable.

```
#Graficar valores extremos de la variable precio del boleto (Fare).
boxplot(Titanic$Fare,
        col = c("51"),
        main = "PRECIO BOLETOS",
        yLab = "$")
```




```
#Visualizamos los valores extremos de la variable precio del boleto (Fare).
boxplot(Titanic$Fare, col = c("51"), main="PRECIO BOLETOS", ylab="$", horizontal = T, plot = F)
```

```
## $stats
##         [,1]
## [1,]  0.0000
## [2,]  7.8958
## [3,] 14.4542
## [4,] 31.2750
## [5,] 65.0000
##
## $n
## [1] 1309
##
## $conf
##          [,1]
## [1,] 13.43322
## [2,] 15.47518
##
## $out
##   [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000
##   [9] 263.0000  77.2875 247.5208  73.5000  77.2875  79.2000  66.6000  69.5500
##  [17]  69.5500 146.5208  69.5500 113.2750  76.2917  90.0000  83.4750  90.0000
##  [25]  79.2000  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500
##  [33]  91.0792 151.5500 247.5208 151.5500 110.8833 108.9000  83.1583 262.3750
##  [41] 164.8667 134.5000  69.5500 135.6333 153.4625 133.6500  66.6000 134.5000
```

```
##  [49] 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
##  [65]  90.0000  78.2667 151.5500  86.5000 108.9000  93.5000 221.7792 106.4250
##  [73]  71.0000 106.4250 110.8833 227.5250  79.6500 110.8833  79.6500  79.2000
##  [81]  78.2667 153.4625  77.9583  69.3000  76.7292  73.5000 113.2750 133.6500
##  [89]  73.5000 512.3292  76.7292 211.3375 110.8833 227.5250 151.5500 227.5250
##  [97] 211.3375 512.3292  78.8500 262.3750  71.0000  86.5000 120.0000  77.9583
## [105] 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  82.2667 262.3750  76.2917 263.0000
## [121] 262.3750 262.3750 263.0000 211.5000 211.5000 221.7792  78.8500 221.7792
## [129]  75.2417 151.5500 262.3750  83.1583 221.7792  83.1583  83.1583 247.5208
## [137]  69.5500 134.5000 227.5250  73.5000 164.8667 211.5000  71.2833  75.2500
## [145] 106.4250 134.5000 136.7792  75.2417 136.7792  82.2667  81.8583 151.5500
## [153]  93.5000 135.6333 146.5208 211.3375  79.2000  69.5500 512.3292  73.5000
## [161]  69.5500  69.5500 134.5000  81.8583 262.3750  93.5000  79.2000 164.8667
## [169] 211.5000  90.0000 108.9000
##
## $group
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## $names
## [1] ""
```

Como podemos observar en el vector stats el ultimo valor que se obtuvo del límite inferior fue de 0 y de 65 para el límite superior y una mediana de 14. Existen una gran cantidad de outliers con valores mayores a 65. No es necesario llevar a cabo ningún tratamiento con los valores extremos de la variable Fare, ya que las personas de adineradas pagaron más por sus boletos para estar ubicados en primera clase y en algunos casos se realizaron pagos exorbitantes. Igualmente, es importante tener en cuenta que la mayoría de pasajeros se encontraban ubicados en la clases 2 y 3 en donde el valor de los boletos no excedía los 20.

#Análisis de los datos

##Selección de los grupos de datos

```
#Agrupación por la clase en la que viajaba el pasajero.
Titanic.1C <- Titanic[Titanic$Pclass == 1,]
Titanic.2C <- Titanic[Titanic$Pclass == 2,]
Titanic.3C <- Titanic[Titanic$Pclass == 3,]

#Agrupación por el sexo del pasajero.
Titanic.Femenino <- Titanic[Titanic$Sex == 0,]
Titanic.Masculino <- Titanic[Titanic$Sex == 1,]

#Agrupación por el edad del pasajero a partir del ciclo de vida.
Titanic.PrimeraInfancia <- Titanic[Titanic$Age <= 5,]
Titanic.Infancia  <- Titanic[Titanic$Age >= 6 & Titanic$Age <= 11,]
Titanic.Adolescencia  <- Titanic[Titanic$Age >= 12 & Titanic$Age <= 18,]
Titanic.Juventud  <- Titanic[Titanic$Age >= 14 & Titanic$Age <= 26,]
Titanic.Adultez  <- Titanic[Titanic$Age >= 27 & Titanic$Age <= 59,]
Titanic.Vejez  <- Titanic[Titanic$Age >= 60,]
```

##Comprobación de la normalidad y homogeneidad de la varianza

Para poder comprobar que las variables numéricas provienen de una distribución normal, emplearemos el test de Shapiro Wilk.

```
#Calcular el p-valor para las variables numéricas del conjunto de datos seleccionado a partir del test
PV <- data.frame(shapiro.test(Titanic$Sex)$p.value,
                 shapiro.test(Titanic$Pclass)$p.value,
                 shapiro.test(Titanic$Age)$p.value,
                 shapiro.test(Titanic$Fare)$p.value)

#Modificar el nombre de las columnas del Data Frame.
PV <- setNames(PV, c("Sex", "Pclass", "Age", "Fare"))

#Validar que el p-valor es superior a un nivel de significancia de 0.05. Si se cumple esta condición se
PV = rbind(PV,data_frame(
  Sex=c(if(PV$Sex<0.05){"NO distribución normal"} else {"SI distribución normal"}),
  Pclass=c(if(PV$Pclass<0.05){"NO distribución normal"} else {"SI distribución normal"}),
  Age=c(if(PV$Age<0.05){"NO distribución normal"} else {"SI distribución normal"}),
  Fare=c(if(PV$Fare<0.05){"NO distribución normal"} else {"SI distribución normal"})))

#Visualizar el p-valor para las variables numéricas y su validación de normalidad.
PV
```

```
##                       Sex                  Pclass                     Age
## 1   1.87029090330203e-47   4.83599691484388e-42    3.3756002826745e-14
## 2 NO distribución normal NO distribución normal NO distribución normal
##                      Fare
## 1   2.40826537485134e-50
## 2 NO distribución normal
```

El test de Shapiro Wilk nos indica que ninguna variable sigue distribución normal, dado que el p-valor es inferior a un nivel de significancia de 0.05.

##Homogeneidad de la varianza

```
#Calcular el p-valor para las variables numéricas del conjunto de datos seleccionado a partir del test
PV <- data.frame(fligner.test(Survived ~ Sex, data = Titanic)$p.value,
                 fligner.test(Survived ~ Pclass, data = Titanic)$p.value,
                 fligner.test(Survived ~ Age, data = Titanic)$p.value,
                 fligner.test(Survived ~ Fare, data = Titanic)$p.value)

#Modificar el nombre de las columnas del Data Frame.
PV <- setNames(PV, c("Sex", "Pclass", "Age", "Fare"))

#Validar que el p-valor es superior a un nivel de significancia de 0.05. Si se cumple esta condición  s
PV = rbind(PV,data_frame(
  Sex=c(if(PV$Sex>0.05){"Varianzas homogéneas"} else {"Varianzas NO homogéneas"}),
  Pclass=c(if(PV$Pclass>0.05){"Varianzas homogéneas"} else {"Varianzas NO homogéneas"}),
  Age=c(if(PV$Age>0.05){"Varianzas homogéneas"} else {"Varianzas NO homogéneas"}),
  Fare=c(if(PV$Fare>0.05){"Varianzas homogéneas"} else {"Varianzas NO homogéneas"})))

#Visualizar el p-valor para las variables numéricas y su validación de homogeneidad de la varianza
PV
```

```
##                       Sex                    Pclass                     Age
## 1    0.000595879736572757    7.84216004993855e-10    0.237667628105039
## 2 Varianzas NO homogéneas Varianzas NO homogéneas Varianzas homogéneas
##                      Fare
```

```
## 1      0.0346153582477336
## 2 Varianzas NO homogéneas
```

##Pruebas estadísticas

###¿Qué variables influyen más en el pasajero haya sobrevivido o no al naufragio?

Como las variables categoricas o cualitativas no siguen una distribución normal emplearemos la correlación no paramétrica de Spearman. Si el coeficiente obtenido es positivo nos indicara que la variable influye en el que el pasajero haya sobrevivido o no al naufragio.

```
#Calcular el coeficiente rho para las variables numéricas del conjunto de datos seleccionado a partir d
RHO<- data.frame(cor.test(Titanic$Survived,Titanic$Sex, method="spearman")$estimate,
                 cor.test(Titanic$Survived,Titanic$Pclass, method="spearman")$estimate,
                 cor.test(Titanic$Survived,Titanic$Age, method="spearman")$estimate,
                 cor.test(Titanic$Survived,Titanic$Fare, method="spearman")$estimate)

#Modificar el nombre de las columnas del Data Frame.
RHO <- setNames(RHO, c("Sex","Pclass", "Age", "Fare"))

#Validar que el coeficiente es positivo. Si se cumple esta condición se considera que variable influye.
RHO = rbind(RHO,data_frame(Sex=c(if(RHO$Sex>0){"SI Influye"} else {"NO Influye"}),
                           Pclass=c(if(RHO$Pclass>0){"SI Influye"} else {"NO Influye"}),
                           Age=c(if(RHO$Age>0){"SI Influye"} else {"NO Influye"}),
                           Fare=c(if(RHO$Fare>0){"SI Influye"} else {"NO Influye"})))

#Visualizar el p-valor para las variables numéricas y su validación de normalidad.
RHO
```

```
##                    Sex            Pclass                  Age              Fare
## rho -0.406513950567393 -0.36776278713577 -0.0443448454869709 0.350653471311607
## 1           NO Influye        NO Influye          NO Influye        SI Influye
```

Como podemos observar el monto que el pasajero pago para obtener su boleto influyó en que halla o no sobrevivido al naufragio.

Contraste de sobrevivencia entre pasajeros de sexo masculino y femenino

####Hipótesis

Nula (H0): La probabilidad de sobrevivir al naufragio de los pasajeros de sexo masculino es igual a la del sexo femenino. (H0:µm = µf)

Alternativa (H1): La probabilidad de sobrevivir al naufragio de los pasajeros de sexo femenino es mayor a la media de sexo masculino (H1:µm < µf)

#####Método

El método propuesto consiste es comprobar si hay diferencias estadísticamente significativas entre la variable sobrevivir (Survived) extraída de dos (2) muestras independientes diferenciadas por la variable sexo (Sex). La primera muestra está compuesta por los pasajeros de sexo masculino (male) y la segunda, por los pasajeros de sexo femenino (female). Con el objetivo de poder visualizar si hay diferencias en las dos (2) muestras.

#####Calculos

```
#Creamos dos (2) muestras para diferenciar los pasajeros de sexo masculino (male) y sexo femenino (fema
Titanic.Male.Survived <- Titanic[Titanic$Sex == 1,]$Survived
Titanic.Female.Survived <- Titanic[Titanic$Sex == 0,]$Survived
```

```r
#Calcular el p-valor para las dos (2) muestras.
t.test(Titanic.Male.Survived, Titanic.Female.Survived, alternative = "less")$p.value
```

```
## [1] 2.372891e-49
```

####Interpretación de los resultados Dado que p-value(2.372891e-49) es menor a un nivel de significancia ( ) de 0.05 se rechaza la hipótesis nula y se dispone de evidencia suficiente para considerar que la probabilidad de sobrevivir al naufragio es mayor en los pasajeros de sexo femenino.

###Regresión lineal

Se llevará a cabo una estimación bajo varios modelos de regresión logística tomando como variable dependiente, sobrevivir o no al naufragio y siendo las variables explicativas, clase en la que se viajaba, sexo, edad y precio del boleto.

```r
#Calculamos varios modelos de regresión logística.

Modelo1 <- lm(formula = Survived ~ Sex + Pclass + Age, data = Titanic)

Modelo2 <- lm(formula = Survived ~  + Pclass + Fare, data = Titanic)

Modelo3 <- lm(formula = Survived ~ Sex + Age + Fare, data = Titanic)

Modelo4 <- lm(formula = Survived ~ Pclass + Age + Fare, data = Titanic)

#Obtenemos los coeficientes de determinación de los modelos.
COE <- data.frame(summary(Modelo1)$r.squared,
                  summary(Modelo2)$r.squared,
                  summary(Modelo3)$r.squared,
                  summary(Modelo4)$r.squared)

#Modificar el nombre de las columnas del Data Frame.
COE <- setNames(COE, c("Modelo1", "Modelo2", "Modelo3", "Modelo4"))

#Visualizar los coeficientes de determinación de los modelos.
COE
```

```
##     Modelo1   Modelo2   Modelo3   Modelo4
## 1 0.3075045 0.1501217 0.2232269 0.2081038
```

Como podemos observar ninguno de los modelos contiene un coeficiente R^2 optimo superior al 60% Sin embargo el que mejor se acerca a la realidad es el modelo 1 con un coeficiente R^2 del 30.7 %.Comprobaremos la validez del modelo 1 realizando la predicción y comparando los valores predecidos con los reales.

```r
#Predecir el estado de sobrevivencia.
V1<-data.frame(Sex=1, Pclass = 1, Age=5)
V2<-data.frame(Sex=1, Pclass = 1, Age=11)
V3<-data.frame(Sex=1, Pclass = 1, Age=18)
V4<-data.frame(Sex=1, Pclass = 1, Age=26)
V5<-data.frame(Sex=1, Pclass = 1, Age=59)
V6<-data.frame(Sex=1, Pclass = 1, Age=60)
V7<-data.frame(Sex=0, Pclass = 1, Age=5)
V8<-data.frame(Sex=0, Pclass = 1, Age=11)
V9<-data.frame(Sex=0, Pclass = 1, Age=18)
V10<-data.frame(Sex=0, Pclass = 1, Age=26)
V11<-data.frame(Sex=0, Pclass = 1, Age=59)
V12<-data.frame(Sex=0, Pclass = 1, Age=60)
```

```
tabla.estimacion <- matrix(c(V1$Sex, V1$Pclass, V1$Age, predict(Modelo1, V1),
                             V2$Sex, V2$Pclass, V2$Age, predict(Modelo1, V2),
                             V3$Sex, V3$Pclass, V3$Age, predict(Modelo1, V3),
                             V4$Sex, V3$Pclass, V4$Age, predict(Modelo1, V4),
                             V5$Sex, V5$Pclass, V5$Age, predict(Modelo1, V5),
                             V6$Sex, V6$Pclass, V5$Age, predict(Modelo1, V6),
                             V7$Sex, V7$Pclass, V7$Age, predict(Modelo1, V7),
                             V8$Sex, V8$Pclass, V8$Age, predict(Modelo1, V8),
                             V9$Sex, V9$Pclass, V9$Age, predict(Modelo1, V9),
                             V10$Sex, V10$Pclass, V10$Age, predict(Modelo1, V10),
                             V11$Sex, V11$Pclass, V11$Age, predict(Modelo1, V11),
                             V12$Sex, V12$Pclass, V12$Age, predict(Modelo1, V12)),
                             ncol = 4, byrow = TRUE)

#Modificar el nombre de las columnas del la tabla de estimación.
colnames(tabla.estimacion) <- c("Sex", "Pclass", "Age", "Survived")

#Visualizamos la tabla de predicciones.
tabla.estimacion
```

```
##       Sex Pclass Age  Survived
##  [1,]   1      1   5 0.7833772
##  [2,]   1      1  11 0.7371759
##  [3,]   1      1  18 0.6832744
##  [4,]   1      1  26 0.6216727
##  [5,]   1      1  59 0.3675658
##  [6,]   1      1  59 0.3598656
##  [7,]   0      1   5 1.1246566
##  [8,]   0      1  11 1.0784554
##  [9,]   0      1  18 1.0245539
## [10,]   0      1  26 0.9629522
## [11,]   0      1  59 0.7088453
## [12,]   0      1  60 0.7011451
```

##Conclusiones

A través del análisis de correlación de las variables cualitativas logramos identificar cuáles son las variables más correlacionadas con el estado de sobrevivencia del pasajero en función de su proximidad con los valores -1 y +1. En conclusión, el monto que el pasajero pagó para obtener su boleto es la variable más relevante para definir si un pasajero sobrevivió al naufragio del Titanic.

El análisis de contraste de sobrevivencia entre pasajeros de sexo masculino y femenino nos permitió identificar que la probabilidad de sobrevivir al naufragio Titanic fue mayor en los pasajeros de sexo femenino. Finalmente, a través de los diferentes modelos de regresión lineal se estableció que el modelo que mejor predecía sobrevivir o no al naufragio es aquel que estaba compuesto por variables explicativas Sex, Pclass, Age.

```
# Procedemos a crear el archivo CSV.
write.table(Titanic, file = "Titanic.csv",row.names=FALSE, na="",col.names=TRUE, sep=",")

# El parametro row.names indica el nombre de cada fila.
# El parametro col.names indica el nombre de las columnas como cabecera en el archivo
# El parametro sep indica el tipo de separador con que registraran los datos en el archivo csv
```

##Contribuciones

Investigación previa: ROSEMBERG ALVAREZ DÍAZ Redacción de las respuestas: ROSEMBERG ALVAREZ DÍAZ Desarrollo código: ROSEMBERG ALVAREZ DÍAZ