Práctica 2: Limpieza y análisis de datos

Diego Labastida

Junio 2022

Contents

Descripción del Dataset

Instalamos y cargamos la librería dplry y ggplot2

A continuación, se describira el juego de datos que se utilizara en sta práctica. El objetivo de este dataset sera resolver preguntas relacionadas con los grupos de sobrevivientes como:

- ¿Cuales fueron los grupos que sobrevivieron al evento en el titanic?
- ¿La clase del boleto tuvo que ver en la supervivencia?
- ¿Que edad tenian los pasajeros del Titanic?
- ¿Los sobrevivientes del Titanic viajaban solos o acompañados?
- ¿La primera clase tuvo privilegios al sobrevivir?

```
genderSubmission <- read.csv('gender_submission.csv',stringsAsFactors = FALSE, sep = ',')
test <- read.csv('test.csv',stringsAsFactors = FALSE, sep = ',')
train <- read.csv('train.csv',stringsAsFactors = FALSE, sep = ',')
```

891 Registros y 12 variables correspondiente a la información del archivo train con información de pasajeros en el Titanic. Los archivos test y gender_submission se complementan y deben unirse posteriortente para obetener 418 Registros con 12 variables.

Finalmente se obtendra un dataset de 1309 Registros y 12 variables.

- **PassengerId: (integer)** Id unico para cada pasajero del Titanic
- **Survived: (integer)** Indica si el pasajero sobrevivio o no (1=sobrevivio, 0=No sobrevivio)
- **Pclass: (integer)** Clase del ticket de los pasajeros 1er, 2da y 3ra clase.
- **Name: (character)** Nombre del Pasajero
- **Sex: (character)** Genero del pasajero (Male / Female) Hombre o Mujer
- **Age: (numeric)** Edad del pasajero
- **SibSp: (integer)** Numero de hermanos a bordo del Titanic
- **Parch: (integer)** Numero de padres abordo del Titanic
- **Ticket: (character)** Numero de tiquete del pasajero
- **Fare: (numeric)** Tarifa del tiquete
- **Cabin: (character)** Numero de cabina
- **Embarked: (character)** Puerto de Embarque

Integración y Selección de los datos

Primero realizaremos una integración de los datos obtenidos en los 3 archivos.

De acuerdo a la variable PassengerId de los juegos de datos ***genderSubmission*** complementamos el dataset ***test***

```
test1 <- merge(test, genderSubmission)
test1 <- test1[order(test1$PassengerId), ]
```

Posteriormente unimos los juegos de datos ***train*** y ***test***

```
dataset <- rbind(train,test1)
filas=dim(dataset)[1]
str(dataset)
```

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

Limpieza de los Datos

Obtenemos una estadistica de valores vacios y NAs de cada columna

```
colSums(is.na(dataset))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         263
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           1           0           0
```

```
colSums(dataset=="")
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0          NA
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0          NA        1014           2
```

Se observan lo siguiente:

- 263 Registros NA en la variable **Age**
- 1 Registro NA para **Fare**
- 1014 Registros vacios para la variable **Cabin**
- 2 Registros vacios para la variable **Embarked**

Asignamos la media para valores vacíos de la variable **Age**.

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

Para el caso de la variable **Fare** la media se debe realizar de acuerdo a la case del ticket, sabiendo que los boletos de cada clase tienen una tarifa completamente distinta por los beneficios que estos pueden tener.

```
mean(dataset$Fare[dataset$Pclass == dataset$Pclass[is.na(dataset$Fare)]],na.rm=T)
```

```
## [1] 13.30289
```

```
dataset$Fare[is.na(dataset$Fare)] <- mean(dataset$Fare[dataset$Pclass == dataset$Pclass[is.na(dataset$Fa
```

Asignamos valor “Desconocido” para los valores vacíos de la variable **Cabin**.

```
dataset$Cabin[dataset$Cabin == ''] <- "Desconocido"
```

Asignamos valor “D” de desconocido para los valores vacíos de la variable **Embarked**.

```
dataset$Embarked[dataset$Embarked == ''] <- "D"
```

Modificamos la variable **Survived** (integer) por una variable Categorica 0=False y 1=True

```
dataset$Survived[dataset$Survived == 0] <- 'False'
dataset$Survived[dataset$Survived == 1] <- 'True'
dataset$Survived <- as.factor(dataset$Survived)
dataset$Pclass <- as.factor(dataset$Pclass)
```

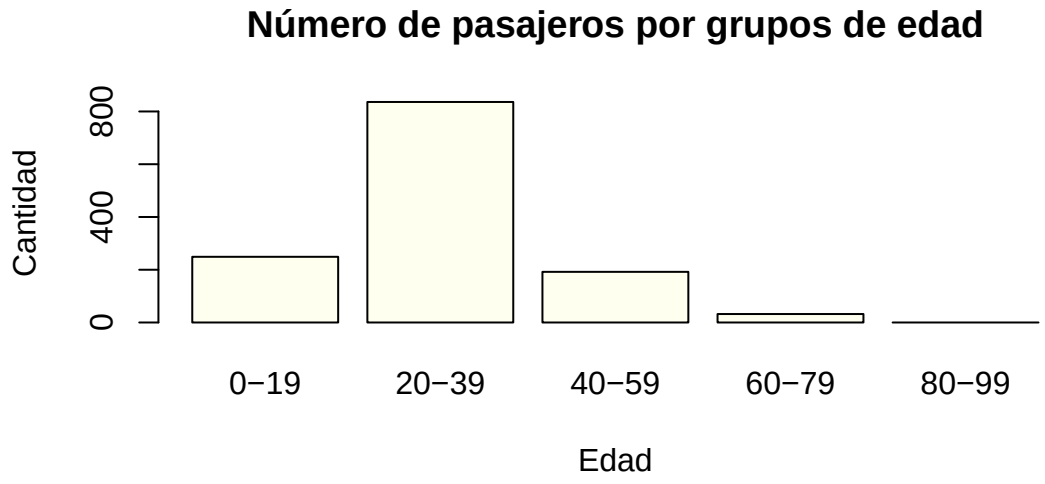
Convertimos la variable**Age** a Integer. Para los valores inferiores a uno se igualaran a la unidad.

```
dataset$Age[dataset$Age < 1 & dataset$Age > 0] <- 1
dataset$Age <- as.integer(dataset$Age)
```

Finalmente hacemos una discretización de las edades de los pasajeros.

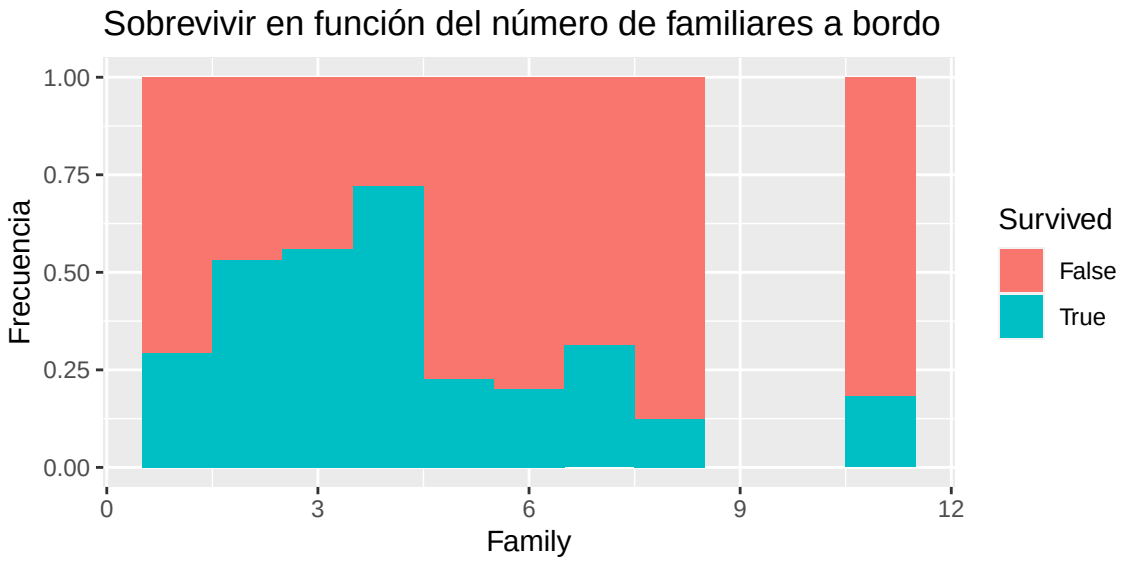
```
dataset["AgeGroup"] <- cut(dataset$Age, breaks = c(0,20,40,60,80,100), labels = c("0-19", "20-39", "40-5
```

```
plot(dataset$AgeGroup,main="Número de pasajeros por grupos de edad",xlab="Edad", ylab="Cantidad",col = '
```



Hacemos la construcción de una variable nueva: **Family**

```
dataset$Family <- dataset$SibSp + dataset$Parch +1;
dataset1<-dataset[1:filas,]
ggplot(data = dataset1[!is.na(dataset[1:filas,]$Family),],aes(x=Family,fill=Survived))+geom_histogram(bi
```

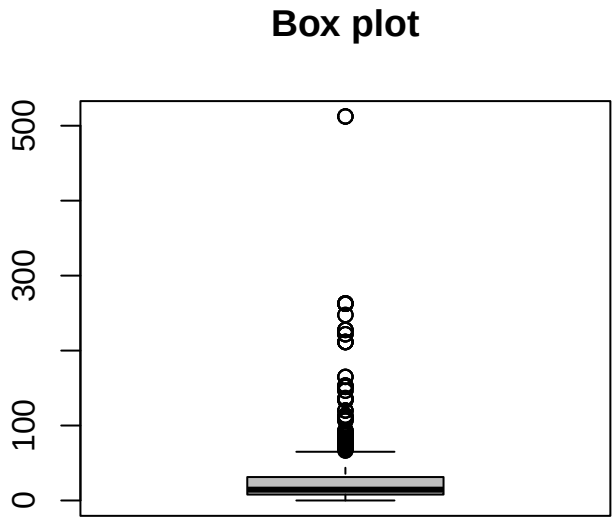


Outliner Fare

Haremos una evaluación de valores atípicos o posibles outliners. La variable **Fare** sera evaluada en su totalidad y posteriormente evaluaremos por la clase del ticket **Pclass**.

Se gráfica todos los valores de la variable ***Fare***

```
boxplot(dataset$Fare,main="Box plot", col="gray")
```



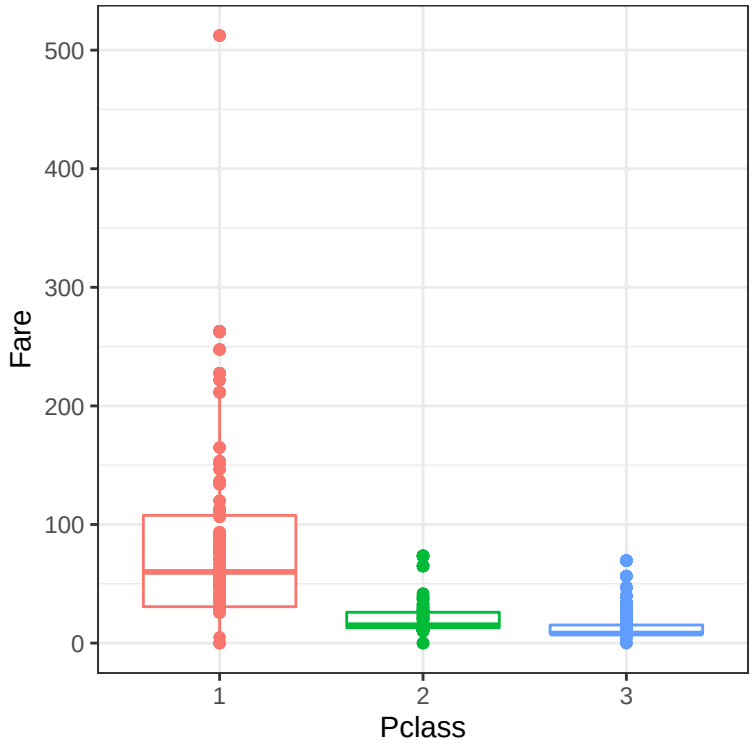
```
unique(boxplot.stats(dataset$Fare)$out)
```

```
##  [1]  71.2833 263.0000 146.5208  82.1708  76.7292  80.0000  83.4750  73.5000
##  [9]  77.2875 247.5208  79.2000  66.6000  69.5500 113.2750  76.2917  90.0000
## [17]  86.5000 512.3292  79.6500 153.4625 135.6333  77.9583  78.8500  91.0792
## [25] 151.5500 110.8833 108.9000  83.1583 262.3750 164.8667 134.5000 133.6500
## [33]  75.2500  69.3000 211.5000 227.5250 120.0000  81.8583  89.1042  78.2667
## [41]  93.5000 221.7792 106.4250  71.0000 211.3375  82.2667  75.2417 136.7792
```

Obtenemos 3 conjuntos de datos con la información de cada clase de Ticket.

```
firstClass <- dataset[dataset$Pclass == 1, ]
secondClass <- dataset[dataset$Pclass == 2, ]
thirdClass <- dataset[dataset$Pclass == 3, ]
```

```
ggplot(data = dataset, aes(x = Pclass, y = Fare, colour = Pclass)) +
  geom_boxplot() +
  geom_point() +
  theme_bw() +
  theme(legend.position = "none")
```



En el primer grafica identificamos 48 posibles outliners pero se considerara evaluar por clase y de esta forma seleccionar los mejores candidatos a ser eliminados.

El segundo gráfico nos muestra la distribución de los outliners para cada Clase.

Se obtienen los valores de la variable **Fare** para la Primera Clase

```
unique(boxplot.stats(firstClass$Fare)$out)
```

```
## [1] 263.0000 247.5208 512.3292 262.3750 227.5250
```

En Primera Clase observamos 5 Outliners como calculo incial se tomaran todos los valores fuera de la caja como outliners y se sustituiran con la media de los datos.

Se obtienen los valores de la variable **Fare** para la Segunda Clase

```
unique(boxplot.stats(secondClass$Fare)$out)
```

```
## [1] 73.5 65.0
```

En la segunda Clase se observa solo 2 valores fuera de la caja, si observamos con atención estan a una distancia considerable por este motivo consideramos ambos valores como Ouliners.

Se obtienen los valores de la variable **Fare** para la Tercera Clase

```
unique(boxplot.stats(thirdClass$Fare)$out)
```

```
## [1] 31.2750 29.1250 31.3875 39.6875 46.9000 27.9000 56.4958 34.3750 69.5500
```

La gráfica para la tercera clase tiene 9 valores candidatos de valores atipicos en este caso se debe mencionar que no todos las tarifas pueden estar erroneas ya que como sabemos las tarifas varian dependiendo del día de su compra, lugar de compra y lugar de abordaje En este caso se hace una selección inicial y de acuerdo a los resultados obtenidos despues de este estudio se pueden modificar para mejorar el resultado final.

```
dataset[dataset$Pclass == 1 & dataset$Fare > 225, ]$Fare <- mean(dataset[dataset$Pclass == 1, ]$Fare)
dataset[dataset$Pclass == 2 & dataset$Fare > 50, ]$Fare <- mean(dataset[dataset$Pclass == 2, ]$Fare)
dataset[dataset$Pclass == 3 & dataset$Fare > 27.5, ]$Fare <- mean(dataset[dataset$Pclass == 3, ]$Fare)
```

Outliners Age

Posteriormente, Se obtenemos los valores de la variable **Age** para identificar si existen valores Outliners

```
unique(boxplot.stats(dataset$Age)$out)
```

```
##  [1]  2 58 55 66 65  1 59 71 70 61 56 62 63 60 64 57 80 74 67 76
```

En el caso de la edad podemos obserbar valores fuera de la caja pero no son necesariamente valores atipicos ya que un humano puede tener desde 1 año hasta mas de 80 años de edad por lo que no podemos considerar que estos valores sean atipicos.

Análisis de los datos.

Comprobación de la Normalidad

Nos proponemos analizar las relaciones entre las diferentes variables del juego de datos para ver si se relacionan y como.

Veamos ahora dos gráficos que nos comparan los atributos Age y Survived. Observamos como el parámetro position="fill" nos da la proporción acumulada de un atributo dentro de otro

Hipótesis

- **H0:** La muestra proviene de una distribución normal.
- **H1:** La muestra no proviene de una distribución normal.

Se realiza la prueba de Shapiro-Wilk para la variable **Age** de todo el dataset y por grupos de genero por la variable **Sex**

```
shapiro.test(dataset$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset$Age
## W = 0.95409, p-value < 2.2e-16
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **La muestra no proviene de una distribución normal**

Se obtiene un subconjuto de datos apartir de la variable **Sex = male**

```
male<-subset(dataset,Sex=="male")
hist(male$Age)
```


```
shapiro.test(male$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male$Age
## W = 0.94157, p-value < 2.2e-16
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **La muestra no proviene de una distribución normal**

Se obtiene un subconjuto de datos apartir de la variable **Sex = female**

```
female<-subset(dataset,Sex=="female")
hist(female$Age)
```


```
shapiro.test(female$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female$Age
## W = 0.96883, p-value = 2.141e-08
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **La muestra no proviene de una distribución normal**

Realizamos pruebas de Normalidad para la variable **Fare**, similar a la variable **Age** hacemos pruebas para la totalidad del dataset y por subgrupos de clase

Hipótesis

- **H0:** La muestra proviene de una distribución normal.
- **H1:** La muestra no proviene de una distribución normal.

Realizamos la prueba de Shapiro-Wilk para la totalidad del dataset con la variable **Fare**

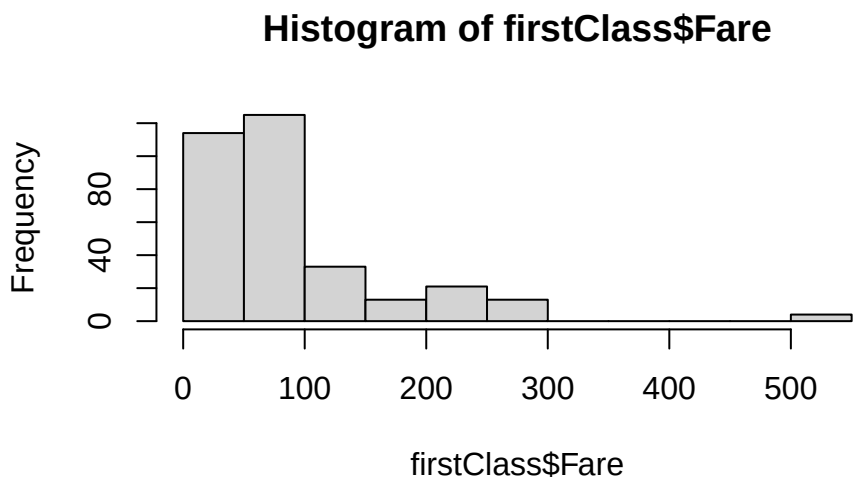
```
shapiro.test(dataset$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset$Fare
## W = 0.6211, p-value < 2.2e-16
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **La muestra no proviene de una distribución normal**

Realizamos la prueba de Shapiro-Wilk para el subconjunto de datos para Primera Clase

```
hist(firstClass$Fare)
```




```
shapiro.test(firstClass$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  firstClass$Fare
## W = 0.74546, p-value < 2.2e-16
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **La muestra no proviene de una distribución normal**

Realizamos la prueba de Shapiro-Wilk para el subconjunto de datos para Segunda Clase

```
hist(secondClass$Fare)
```


```
shapiro.test(secondClass$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  secondClass$Fare
## W = 0.7751, p-value < 2.2e-16
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **La muestra no proviene de una distribución normal**

Realizamos la prueba de Shapiro-Wilk para el subconjunto de datos para Tercera Clase

```
hist(thirdClass$Fare)
```

```r
shapiro.test(thirdClass$Fare)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  thirdClass$Fare
## W = 0.59195, p-value < 2.2e-16
```

Homogeneidad de la varianza

Hipótesis

- **H0:** Las varianzas son homogeneas
- **H1:** Las varianzas no son homogeneas

Ya que conocemos que las datos no son normales usaremos: *Test de Leven* y *Fligner-Killen.*

```r
fligner.test(Fare ~ Sex, dataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Fare by Sex
## Fligner-Killeen:med chi-squared = 95.04, df = 1, p-value < 2.2e-16
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **Las varianzas no son homogeneas**

```r
fligner.test(Age ~ Sex, dataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Age by Sex
## Fligner-Killeen:med chi-squared = 4.104, df = 1, p-value = 0.04278
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **Las varianzas no son homogeneas**

```r
fligner.test(Family ~ Sex, dataset)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  Family by Sex
## Fligner-Killeen:med chi-squared = 83.265, df = 1, p-value < 2.2e-16
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **Las varianzas no son homogeneas**

```
leveneTest(y = dataset$Family, group = dataset$Pclass, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value  Pr(>F)
## group    2  2.5351 0.07964 .
##       1306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La evidencia de acuerdo al p-valor nos dice que no podemos rechazar la hipotesis nula. Por lo tanto **Las varianzas son homogeneas**

```
leveneTest(y = dataset$Fare, group = dataset$AgeGroup, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##         Df F value    Pr(>F)
## group    3  13.312 1.461e-08 ***
##       1305
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La evidencia de acuerdo al p-valor nos dice que podemos rechazar la hipotesis 0. Por lo tanto **Las varianzas no son homogeneas**

Comprobación de Correlación de variables

Ya que nuestros datos no siguen una distribución normal para las variables anteriores, se utilizara el coeficiente de correlación de **Spearman**

Hipótesis

- **H0:** La variables tienen una correlación.
- **H1:** La variables no tienen una correlación.

Vamos a probar si hay una correlación entre la edad del pasajero y el que pagó por el viaje

```
# https://cran.r-project.org/web/packages/tidyverse/index.html
cor.test(x = dataset$Age, y = dataset$Fare, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  dataset$Age and dataset$Fare
## S = 297264437, p-value = 7.314e-14
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2048016
```

```
plotCo1 <- ggplot(data = dataset, aes(x = Age, y = log(Fare))) + geom_point(color = "gray30") + geom_sm
```

Cómo podemos observar no parece haber correlación lineal entre la edad del pasajero y el precio del billete. El diagrama de dispersión tampoco apunta a ningún tipo de relación no lineal evidente.

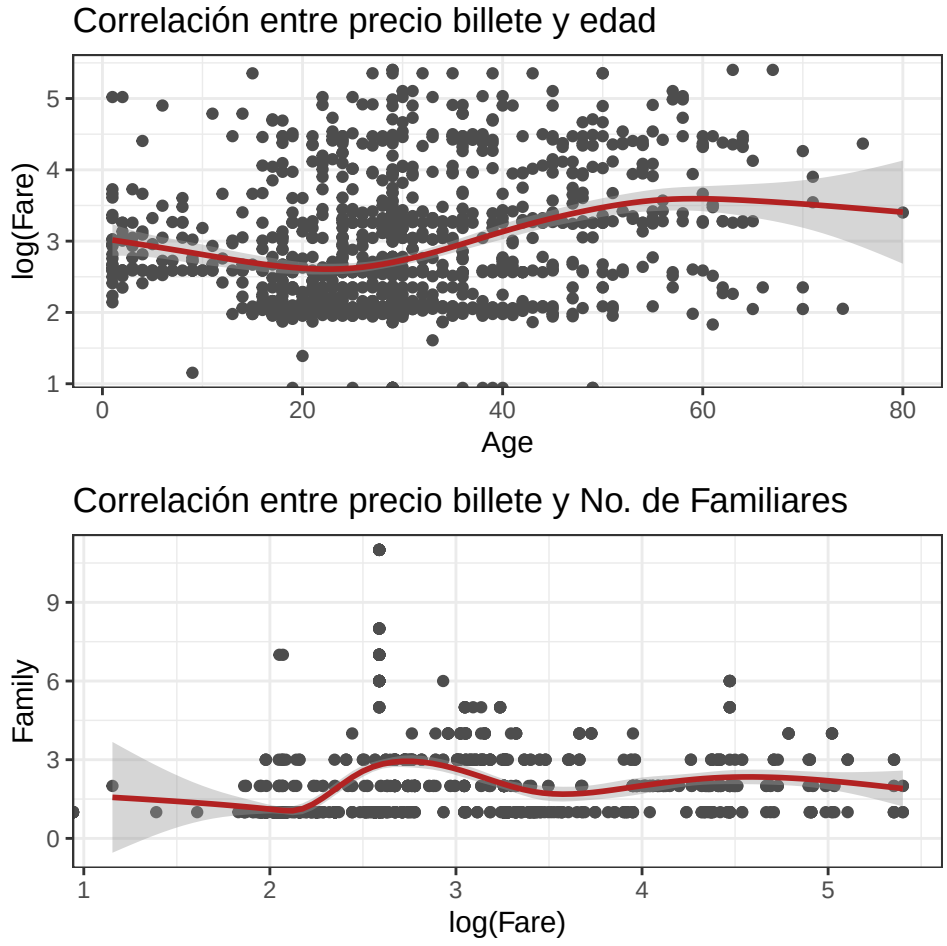
Vamos a probar si hay una correlación entre el que pagó por el viaje y el número en la familia

```
# https://cran.r-project.org/web/packages/tidyverse/index.html
cor.test(x = dataset$Fare, y = dataset$Family, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  dataset$Fare and dataset$Family
## S = 198980561, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4677162
```

```
plotCo2 <- ggplot(data = dataset, aes(x = log(Fare), y = Family)) + geom_point(color = "gray30") + geom_

grid.arrange(plotCo1, plotCo2)
```



Cómo podemos observar no parece haber correlación lineal entre el precio del billete y el número en la familia. El diagrama de dispersión tampoco apunta a ningún tipo de relación no lineal evidente.

Correlación entre variables Categoricas

Hipótesis

- **H0:** No existe asociación entre las variables.
- **H1:** Hay asociación entre las variables.

```
attach(dataset)
chisq.test(table(Pclass,Survived))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Pclass, Survived)
## X-squared = 91.724, df = 2, p-value < 2.2e-16
```

```
chisq.test(table(Sex,Survived))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(Sex, Survived)
## X-squared = 617.31, df = 1, p-value < 2.2e-16
```

```
chisq.test(table(Age,Survived))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Age, Survived)
## X-squared = 117.15, df = 71, p-value = 0.0004695
```

```
chisq.test(table(Family,Survived))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Family, Survived)
## X-squared = 101.97, df = 8, p-value < 2.2e-16
```

Comprobando el valor de PiCuadrado podemos observar una correlación entre las variables anteriores en el siguiente orden, teniendo la mayor correlación la primer variable: Sex, Age, Family y Pclass

Pruebas de contraste de hipótesis

Hipótesis

¿La proporcuón de personas con 20-39 años es superior a las personas con 0-19 años?

- **H0:** PA = PB
- **H1:** PA > PB

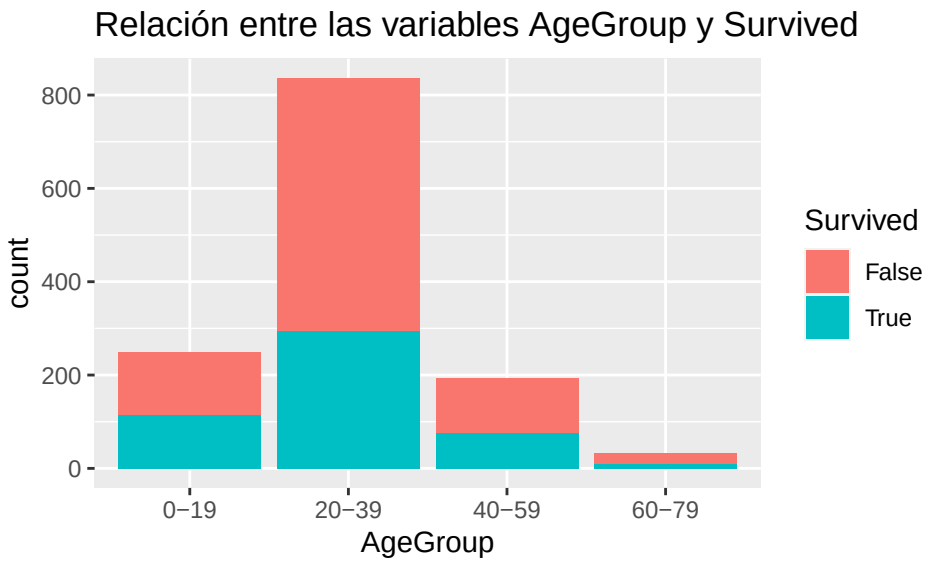
```
x1 <- dataset[dataset$AgeGroup == '20-39', ]
x2 <- dataset[dataset$AgeGroup == '0-19', ]
n1 <- length( x1$Survived )
n2 <- length( x2$Survived )
p1 <- sum(x1$Survived == 'True')/n1;
p2 <- sum(x2$Survived == 'True')/n2;
success<-c( p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  success out of nn
## X-squared = 8.9995, df = 1, p-value = 0.9986
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.1635788  1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.3528708 0.4578313
```

El valor p-valor es superior al numero alpha cayendo en la zona de aceptación de la hipótesis nula. De manera que la proporción de personas con edad de 20-39 años no es superior a las personas con 0-19 años de edad.

A continuación se observa la distribución de sobrevivientes de acuerdo a las variables **Age**, **GroupAge** VS **Survived**

```
ggplot(data=dataset[1:filas,],aes(x=AgeGroup,
          fill=Survived))+geom_bar()+ggtitle("Relación entre las variables AgeGroup y Survived")
```



Hipótesis

¿Las personas con menos de 2 familiares sobrevivieron al Titanic?

- **H0:** media = PB
- **H1:** media < PB

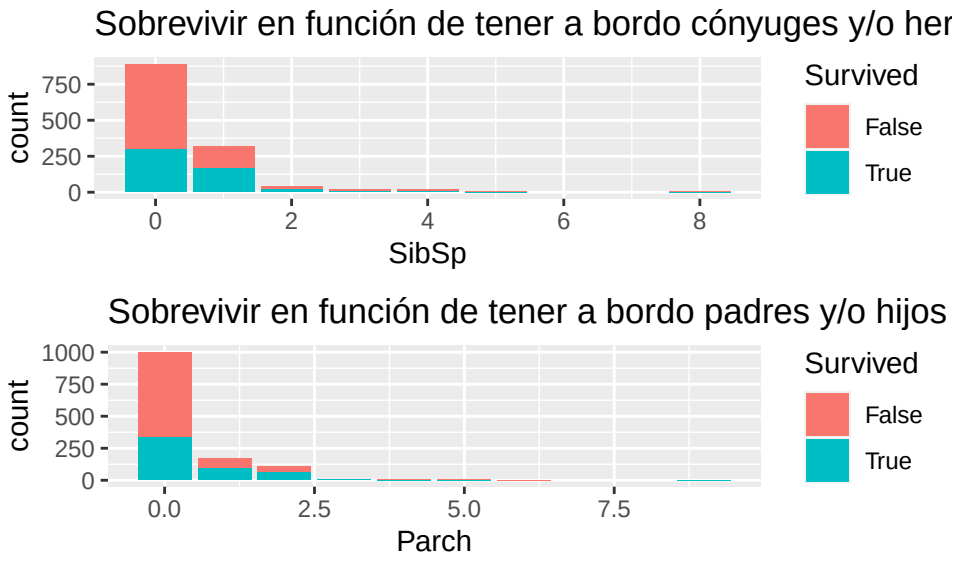
```
t.test( dataset$Family, alternative="less", mu=2)
```

```
##
##  One Sample t-test
##
## data:  dataset$Family
## t = -2.6529, df = 1308, p-value = 0.004039
## alternative hypothesis: true mean is less than 2
## 95 percent confidence interval:
##      -Inf 1.955929
## sample estimates:
## mean of x
##  1.883881
```

El p-valor es inferior a nuestro numero alpha por lo que se rechaza la hipotesis nula (Ho). Por lo tanto, las personas con menos de 2 familiares sobrevivieron a el Titanic.

Acontinuación se gráfica la relación entre las variables **SibSp** (# of siblings / supouses aboard the Titanic) y **Parch**(# of parents / children aboard the Titanic)

```
plot1 <- ggplot(data = dataset[1:filas,],aes(x=SibSp,fill=Survived))+geom_bar()+ggtitle("Sobrevivir en
plot2 <- ggplot(data = dataset[1:filas,],aes(x=Parch,fill=Survived))+geom_bar()+ggtitle("Sobrevivir en
grid.arrange(plot1, plot2)
```



Comprobamos que nuestro test de hipotesis esta en lo correcto ya que los sobrevienvtes con 1 o ningun familiara sobrevivieron en su mayoria.

Hipótesis

¿LA proporcion de pasajeros de 1ra Clase que sobrevivio es superior a un 50%?

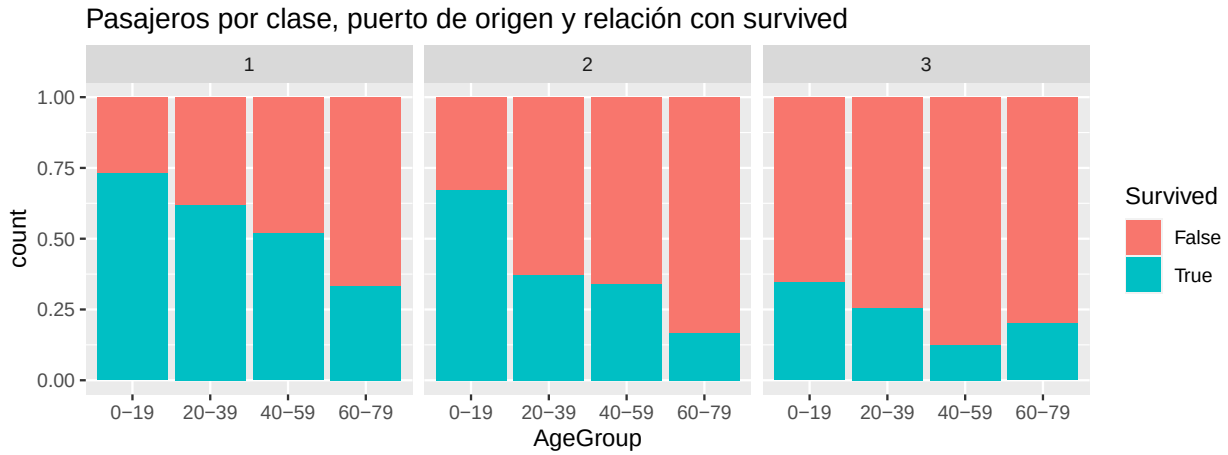
- **H0:** p = 0.5
- **H1:** p > 0.5

```
prop.test(x=sum(dataset$Pclass == 1 & dataset$Survived == 'True'), n=sum(dataset$Pclass == 1), p=0.5, al
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  sum(dataset$Pclass == 1 & dataset$Survived == "True") out of sum(dataset$Pclass == 1), null pr
## X-squared = 7.4334, df = 1, p-value = 0.003201
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5301737 1.0000000
## sample estimates:
##         p
## 0.5758514
```

El p-valor es inferior a alpha por lo que se puede rechazar la hipotesis nula y podemos decir que la proporción de pasajeros en 1ra clase que se salvaron fue de mas del 50%.

```
ggplot(data = dataset[1:filas,],aes(x=AgeGroup,fill=Survived))+geom_bar(position="fill")+facet_wrap(~Pcl
```



Hipótesis

¿La proporción de pasajeros de 1ra clase sobrevivientes es superior a los de 2da y 3ra clase?

- **H0:** PA = PB+PC
- **H1:** PA > PB+PC

```
x1 <- dataset[dataset$Pclass == 1, ]
x2 <- dataset[dataset$Pclass != 1, ]
n1 <- length( x1$Survived )
n2 <- length( x2$Survived )
p1 <- sum(x1$Survived == 'True')/n1;
p2 <- sum(x2$Survived == 'True')/n2;
success<-c( p1*n1, p2*n2)
nn<-c(n1,n2)
prop.test(success, nn, alternative="greater", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  success out of nn
## X-squared = 71.883, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.2121433 1.0000000
## sample estimates:
##    prop 1    prop 2
## 0.5758514 0.3123732
```

El p-valor es inferior a Alpha por lo que podemos rechaazar la hipotesis nula y podemos decir que la proporcion de sobrevivientes de pasajeros en 1ra Clase es superior a la proporción de sobrevivientes en clase 2da y 3ra.

Conclusiones

Usando la Estadistica Inferencial podemos responder las preguntas iniciales que nos planteamos como objetivo a resolver.

- ¿Cuales fueron los grupos que sobrevivieron al evento en el titanic? De acuerdo al estudio realizado los grupos con mayores posibilidades de supervivencia son principalmente el Sexo Femenino y los pasajeros con Boleto de 1era clase. Aunque es interesante decir que tambien el numero de Familiares es un indicador de mayor supervivencia, aunque para este estudio se evalua la presencia de hijos, hermanos o padres, es posible destacar que una persona que viajaba solo o en pareja tenia mayor probabilidad de sobrevivir.
- ¿La clase del boleto tuvo que ver en la supervivencia?
- ¿La primera clase tuvo privilegios al sobrevivir? Como se menciona anteriormente 1era clase tuvo mayores oportunidades de supervivencia a comparación de las otras Clases y eso lo pudimos notar en la prueba de hipotsis realiza al comparar las proporciones de sobrevivientes para ambos grupos.
- ¿Que edad tenian los pasajeros del Titanic? En su mayoria eran pasajeros menores a 40 años, en este caso no se utiliza una Estadistica Inferencial pero si DEscriptiva al observar la distribución de edades.
- ¿Los sobrevivientes del Titanic viajaban solos o acompañados? Los sobrevivientes del titanic en su mayoria eran personas sin familia en el titanic, aunque se puede destacar que nuestro dataset no habla de parejas casadas o novios, solo tenemos valores de familiares de sangre hijos, hermanos, padres.

```
kable(resultados)
```

Contribuciones.	Integrante
Investigación Previa	Diego Labastida
Redacción de las respuestas	Diego Labastida
Desarrollo de Código	Diego Labastida