Modelamiento Estadístico y Sistemas Recomendadores: Foro 1

Patricio Águila Márquez

I. Considere los datos 'hours_peer_week.csv', que contiene las horas que trabaja un grupo de trabajadores de EE.UU. a la semana.

1) Cargue el conjunto de datos en la sesión de trabajo de R usando la función read.table.

```
#Para leer el set de datos, se usará la función 'read.csv'
datos <- read.csv("../Archivos R/hour_per_week.csv",header=TRUE)
```

2) Calcule en forma manual el puntaje Z para las horas de trabajo semanal.

```
#Cálculo de la media
mean_o <- mean(datos$hour_per_week)
#Cálculo de la desviación estándar
desv_o <- sd(datos$hour_per_week)
#Valor de la media, redondeado a dos decimales
round(mean_o,2)
```

```
## [1] 40.44
```

```
#Valor de la desviación estándar, redondeado a dos decimales
round(desv_o,2)
```

```
## [1] 12.35
```

```
#Cálculo del puntaje Z: todas las observaciones (32.561) son normalizadas
puntaje_z <- (datos$hour_per_week-mean_o)/desv_o
#Cálculo de la media normalizada
mean_z <- mean(puntaje_z)
#Cálculo de la desviación normalizada
desv_z <- sd(puntaje_z)
#Valor de la media normalizada
round(mean_z,2)
```

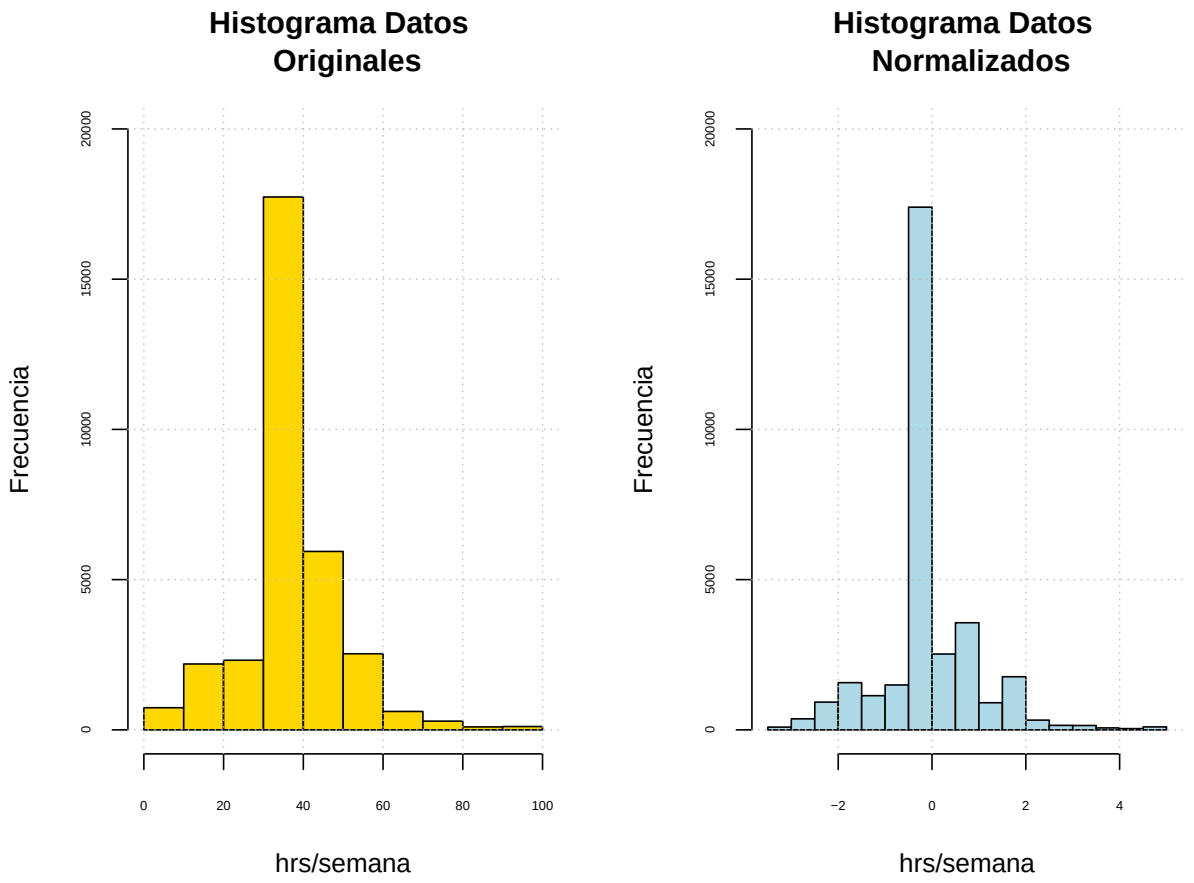
```
## [1] 0
```

```
#Valor de la desviación normalizada
round(desv_z,2)
```

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

3) Construya un histograma de los datos originales y los datos estandarizados. Describa las características principales de los datos, comentando en el foro sobre la simetría y uni- o multi-modalidad de la distribución de los datos.

```
#Gráficos de histograma mostrados en una fila y dos columnas
par(mfrow=c(1,2))
#Histograma de los datos originales, mostrados de 10 en 10
hist(datos$hour_per_week,10,
     main = "Histograma Datos \n Originales",
     ylab = "Frecuencia",
     xlab = "hrs/semana",
     ylim = c(0, 20000),
     col="gold",
     cex.axis=0.5)
abline(h=seq(0,20000,5000), v = seq(0,100,20), lty="dotted", col = "grey")
#Hisotrama de los datos normalizados
hist(puntaje_z,
     main = "Histograma Datos \n Normalizados",
     ylab = "Frecuencia",
     xlab = "hrs/semana",
     ylim = c(0, 20000),
     col = "light blue",
     cex.axis=0.5)
abline(h=seq(0,20000,5000), v = seq(-2,4,2), lty="dotted", col = "grey")
```



```r
#Construcción de indicadores adicionales para respoder a la pregunta 3:

#Función 'summary'
summary(datos)
```

```
##  hour_per_week
##  Min.   : 1.00
##  1st Qu.:40.00
##  Median :40.00
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

```r
#Cálculo manual de la moda
moda <- function(d){
  names(which(table(d)==max(table(d))))
}
#Valor de la moda
moda_o <- as.numeric(moda(datos$hour_per_week))
moda_o
```

```
## [1] 40
```

```r
#Cálculo manual del coeficiente de asimetría de Pearson:
coef_pearson <- (mean_o - moda_o)/desv_o
round(coef_pearson,2)
```

```
## [1] 0.04
```

Respuesta a pregunta 3:

- Los datos de las horas trabajadas por semana presentan una distribución unimodal y levemente asimétrica hacia la derecha.
  - Unimodal, dado que tienen un solo peak, en las 40 horas.
  - Levemente asimétrica hacia la derecha, debido a su coeficiente de Pearson (*), el cual es positivo, lo que indicaría que hay más valores separados de la media hacia la derecha.
- De la representación gráfica, se puede determinar que el 50% de los datos está concentrado entre las 40 y 45 horas.

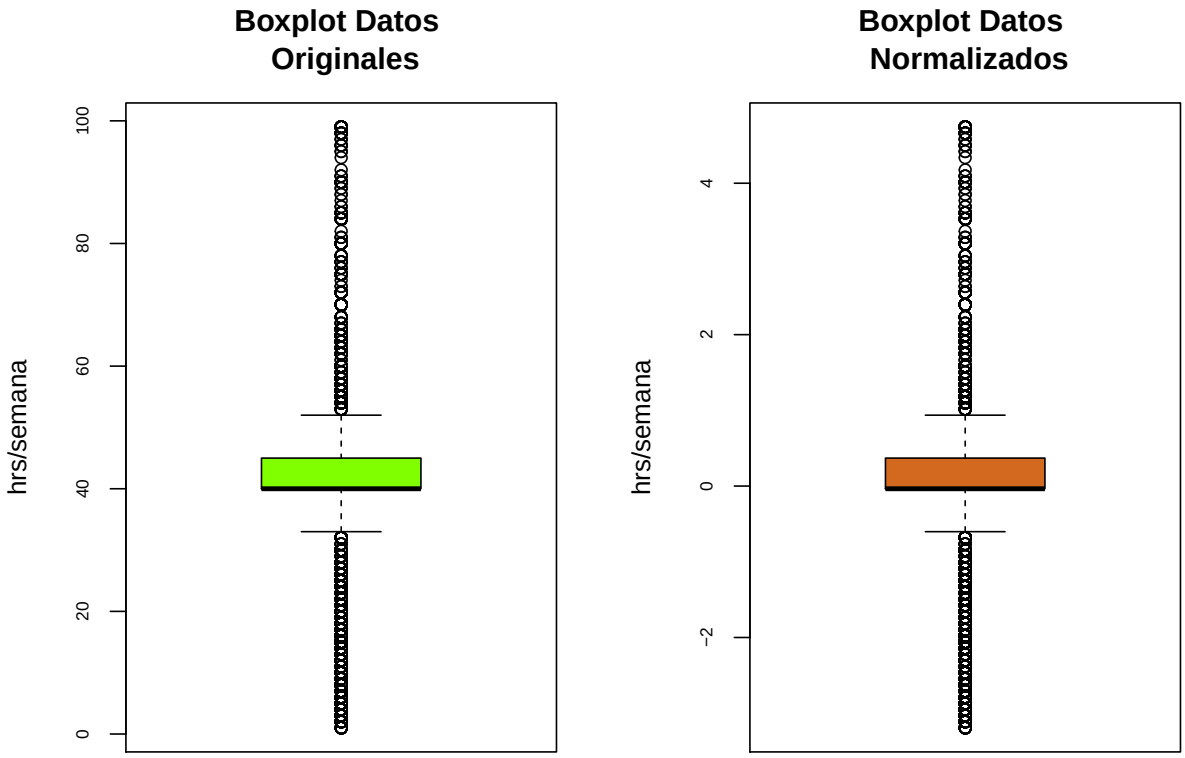
(*) [Pearson's Coefficient of Skewness,

https://www.statisticshowto.com/pearsons-coefficient-of-skewness/,

Interpretation]

4) Construya un boxplot de los datos originales y los datos estandarizados. Comente sus resultados en el foro. ¿Existe evidencia de la presencia de "outliers"? Justifique su respuesta en el foro.

```
#Distribución de gráficos en una fila y dos columnas
par(mfrow=c(1,2))
#Boxblot datos originales
boxplot(datos$hour_per_week,
        main = "Boxplot Datos \n Originales",
        ylab = "hrs/semana",
        col = "chartreuse",
        cex.axis=0.7)
#Boxplot datos normalizados
boxplot(puntaje_z,
        main = "Boxplot Datos \n Normalizados",
        ylab = "hrs/semana",
        col = "chocolate",
        cex.axis=0.7)
```

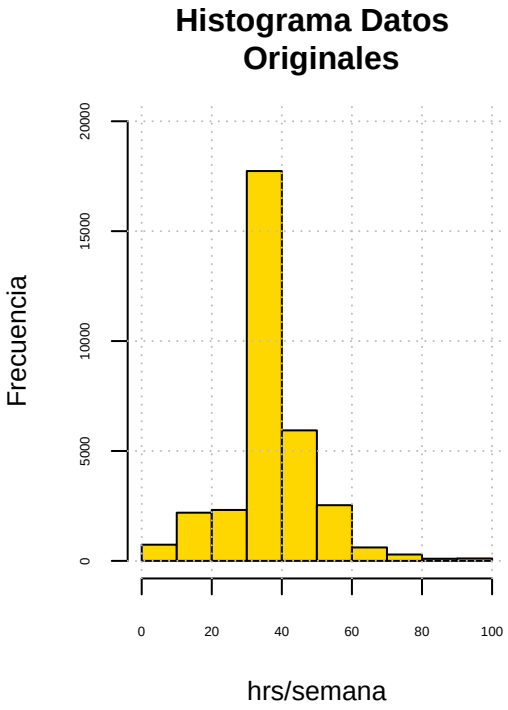


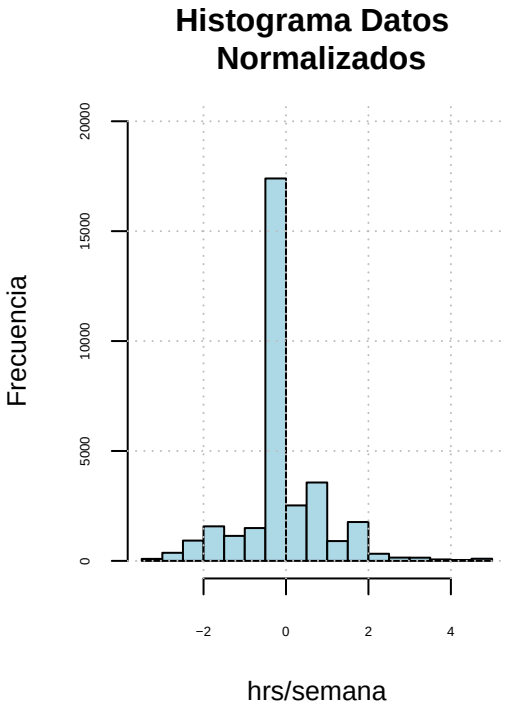
Respuesta a pregunta 4:

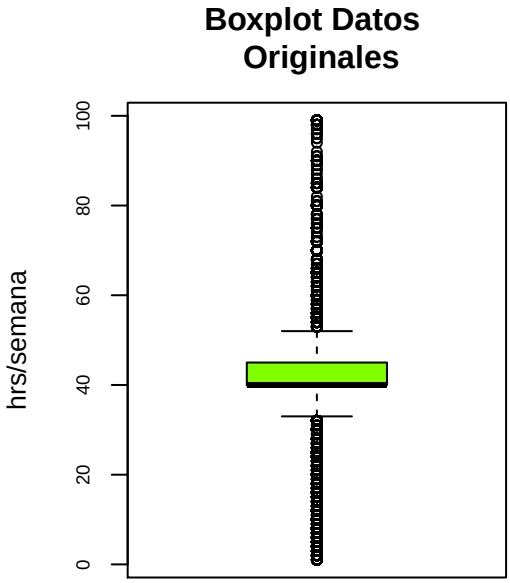
- El boxplot muestra la presencia de valores outliers. Existe mucha dispersión de datos, principalmente hacia la derecha.
- Sería de gran ayuda conocer variables como el departamento, cargo, ubicación geográfica, edad y sexo de los trabajadores, los que podrían arrojar luces sobre las razones de los valores atípicos.

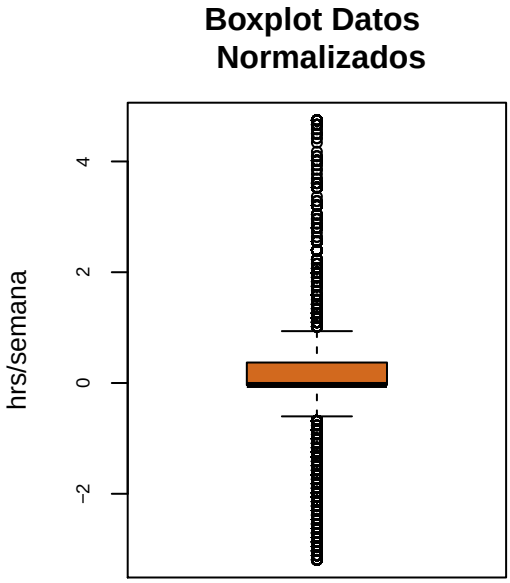
5) Repita los pasos 3) y 4) usando la función scale.

```
#Se agrupan los 4 gráficos en dos filas y dos columnas
par(mfrow=c(2,2))
#Histograma de los datos originales
hist(datos$hour_per_week, 10,
     main = "Histograma Datos \n Originales",
     ylab = "Frecuencia",
     xlab = "hrs/semana",
     ylim = c(0, 20000),
     col="gold",
     cex.axis=0.5)
abline(h=seq(0,20000,5000), v = seq(0,100,20), lty="dotted", col = "grey")
#Hisotrama de los datos normalizados
hist(scale(datos$hour_per_week),
     main = "Histograma Datos \n Normalizados",
     ylab = "Frecuencia",
     xlab = "hrs/semana",
     ylim = c(0, 20000),
     col = "light blue",
     cex.axis=0.5)
abline(h=seq(0,20000,5000), v = seq(-2,4,2), lty="dotted", col = "grey")
#Boxblot datos originales
boxplot(datos$hour_per_week,
        main = "Boxplot Datos \n Originales",
        ylab = "hrs/semana",
        col = "chartreuse",
        cex.axis=0.7)
#Boxplot datos normalizados
boxplot(scale(datos$hour_per_week),
        main = "Boxplot Datos \n Normalizados",
        ylab = "hrs/semana",
        col = "chocolate",
        cex.axis=0.7)
```


Histograma Datos
Originales
Frecuencia
0
5000
10000
15000
20000
0
20
40
60
80
100
hrs/semana


Histograma Datos
Normalizados
Frecuencia
0
5000
10000
15000
20000
−2
0
2
4
hrs/semana


Boxplot Datos
Originales
hrs/semana
0
20
40
60
80
100


Boxplot Datos
Normalizados
hrs/semana
−2
0
2
4

II. Considere los datos 'titanic.csv', sobre la tragedia del Titanic, cuya descripción se muestra en la siguiente tabla.

Variable	Descripción
passengerId	Identificador de pasajero
Survived	Variable que indica 1 si el pasajero sobrevivió y 0 si no.
Pclass	Clase del pasajero (1=primera clase, 2=segunda clase, 3=tercera clase)
Name	Nombre del pasajero
Sex	Género del pasajero
Age	Edad del pasajero
Sibsp	Número de hermanos o cónyuges a bordo
Parch	Número de padres o hermanos a bordo
Ticket	Número de ticket
Fare	Precio del ticket (en moneda local)
embarked	Puerto de embarque (C = Cherbourg; Q = Queenstown; S = Southampton)

Realice las siguientes actividades:

1) Cargue el conjunto de datos en la sesión de trabajo de R usando la función read.table.

```
#Para leer el set de datos, se usará la función 'read.csv'
titanic_data <- read.csv("../Archivos R/titanic.csv",header=TRUE)
```

2) Usando la función summary(), obtenga estadísticos descriptivos de las variables y discuta los resultados en el foro.

```
#Lectura de la data sin procesar
summary(titanic_data)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare          Embarked            Title
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
```

```
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```r
#Detalle del tipo de variables sin procesar
str(titanic_data)
```

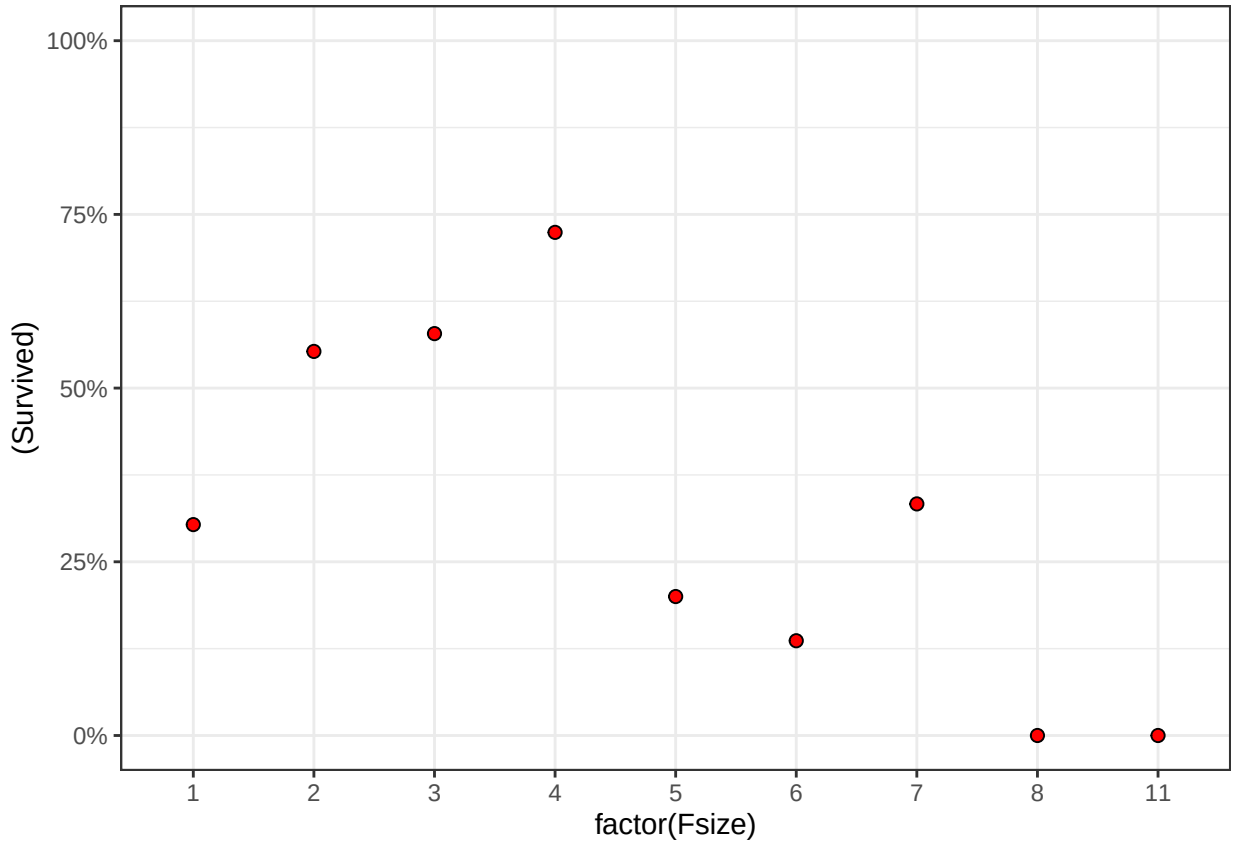
```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
##  $ Title      : chr  "Mr" "Mrs" "Miss" "Mrs" ...
```

3) Cree una variable que indique el tamaño total de la familia del pasajero (incluyéndose él mismo).

```
#El tamaño de la familia se obtiene sumando las columnas 'SibSp', 'Parch',
#incluyendo al pasajero
titanic_data$Fsize <- titanic_data$SibSp + titanic_data$Parch + 1
```

4) Grafique la relación entre la tasa de sobrevivientes y el tamaño de la familia.

```
#Se cargan las librerías para procesar los datos
library(ggplot2)
library(ggthemes)
library(scales)

#Creación de la gráfica "tamaño familia vs % supervivencia"
ggplot(titanic_data, aes(factor(Fsize), (Survived))) +
        stat_summary(fun=mean, geom="point", shape=21, fill="red", size=2) +
        scale_y_continuous(labels=percent_format(), limits=c(0,1)) +
        theme_bw()
```



5) En base a lo observado en el punto anterior, proponga e implemente la discretización del tamaño de la familia. Justifique su decisión en el foro.

```
#Para poder fundamentar la respuesta a la pregunta 5, se realizan previamente,
#los siguientes cálculos:

#Cantidad de pasajeros por tamaño de familia
table(titanic_data$Fsize)
```

```
##
##   1   2   3   4   5   6   7   8  11
## 537 161 102  29  15  22  12   6   7
```

```
#Porcentaje de supervivencia por tamaño de familia
round(prop.table(table(titanic_data$Survived,
                       titanic_data$Fsize),margin = 2),2)
```

```
##
##        1    2    3    4    5    6    7    8   11
##   0 0.70 0.45 0.42 0.28 0.80 0.86 0.67 1.00 1.00
##   1 0.30 0.55 0.58 0.72 0.20 0.14 0.33 0.00 0.00
```

```
#Se propone la siguiente discretización:
#Familias de tamaño 1 (1), de 2 a 4 (2), de 5 a 7 (3) y mayor a 7 (4).

#Familia de 1 integrante
titanic_data$FsizeD[titanic_data$Fsize == 1] <- 1
#Familia de 2 a 4 integrantes
titanic_data$FsizeD[titanic_data$Fsize < 5 & titanic_data$Fsize > 1] <- 2
#Familia de 5 a 7 integrantes
titanic_data$FsizeD[titanic_data$Fsize < 8 & titanic_data$Fsize > 4] <- 3
#Familia de 8 o más integrantes
titanic_data$FsizeD[titanic_data$Fsize > 7] <- 4

#Cantidad de pasajeros por tamaño de familia discretizada
table(titanic_data$FsizeD)
```

```
##
##   1   2   3   4
## 537 292  49  13
```

```
##Porcentaje de supervivencia por tamaño de familia discretizada
round(prop.table(table(titanic_data$Survived,
                       titanic_data$FsizeD),margin = 2),2)
```

```
##
##        1    2    3    4
##   0 0.70 0.42 0.80 1.00
##   1 0.30 0.58 0.20 0.00
```

6) Identifique los pasajeros con datos faltantes para la variable embarked y age usando la función is.na(). ¿Qué tipo de mecanismo de generación de datos faltantes podría ser válido en cada caso? Justifique su respuesta en el foro.

```
#Creación de una variable que filtre los valores NA en la columna 'Embarked'
Embarked_NA <- titanic_data[is.na(titanic_data$Embarked),]
#Listado de pasajeros con valores NA en la columna 'Embarked'
Embarked_NA
```

```
##     PassengerId Survived Pclass                                      Name
## 62           62        1      1                      Icard, Miss. Amelie
## 830         830        1      1 Stone, Mrs. George Nelson (Martha Evelyn)
##        Sex Age SibSp Parch Ticket Fare Embarked Title Fsize FsizeD
## 62  female  38     0     0 113572   80     <NA>  Miss     1      1
## 830 female  62     0     0 113572   80     <NA>   Mrs     1      1
```

```
#Creación de una variable que filtre los valores NA en la columna 'Age'
Age_NA <- titanic_data[is.na(titanic_data$Age),]
#Listado de los 10 primeros pasajeros con valores NA en la columna 'Age'
head(Age_NA,10)
```

```
##    PassengerId Survived Pclass                                      Name
## 6            6        0      3                          Moran, Mr. James
## 18          18        1      2                  Williams, Mr. Charles Eugene
## 20          20        1      3                       Masselmani, Mrs. Fatima
## 27          27        0      3                       Emir, Mr. Farred Chehab
## 29          29        1      3                  O'Dwyer, Miss. Ellen "Nellie"
## 30          30        0      3                            Todoroff, Mr. Lalio
## 32          32        1      1 Spencer, Mrs. William Augustus (Marie Eugenie)
## 33          33        1      3                       Glynn, Miss. Mary Agatha
## 37          37        1      3                              Mamee, Mr. Hanna
## 43          43        0      3                          Kraeff, Mr. Theodor
##       Sex Age SibSp Parch   Ticket     Fare Embarked Title Fsize FsizeD
## 6    male  NA     0     0   330877   8.4583        Q    Mr     1      1
## 18   male  NA     0     0   244373  13.0000        S    Mr     1      1
## 20 female  NA     0     0     2649   7.2250        C   Mrs     1      1
## 27   male  NA     0     0     2631   7.2250        C    Mr     1      1
## 29 female  NA     0     0   330959   7.8792        Q  Miss     1      1
## 30   male  NA     0     0   349216   7.8958        S    Mr     1      1
## 32 female  NA     1     0 PC 17569 146.5208        C   Mrs     2      2
## 33 female  NA     0     0   335677   7.7500        Q  Miss     1      1
## 37   male  NA     0     0     2677   7.2292        C    Mr     1      1
## 43   male  NA     0     0   349253   7.8958        C    Mr     1      1
```

```
#Total de pasajeros con valores NA en la columna 'Age'
nrow(Age_NA)
```

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

7) Genere un conjunto de datos completos al imputar los valores faltantes de la variable embarked por la del puerto "C" y los valores faltantes de la variable edad por el promedio de edad de los datos observados.

```
#Reemplazar datos 'NA' en columna 'Embarked' por puerto 'C'
titanic_data$Embarked[Embarked_NA$PassengerId] <- 'C'
#Reemplazar datos 'NA' en columna 'Age' por el valor de la media de 'Age'
titanic_data$Age[Age_NA$PassengerId] <- mean(na.omit(titanic_data$Age))
```

```
#Para la visualización final de los datos, se convierten a factor las columnas:
#Survived, Pclass, Sex, Embarked, Title, FsizeD
cols<-c('Survived','Pclass','Sex','Embarked','Title','FsizeD',
        'Fsize','SibSp','Parch')
for (i in cols){
  titanic_data[,i] <- as.factor(titanic_data[,i])
}

#Resumen de los datos procesados
summary(titanic_data)
```

```
##   PassengerId    Survived Pclass      Name              Sex
##  Min.   :  1.0   0:549    1:216   Length:891         female:314
##  1st Qu.:223.5   1:342    2:184   Class :character   male  :577
##  Median :446.0            3:491   Mode  :character
##  Mean   :446.0
##  3rd Qu.:668.5
##  Max.   :891.0
##
##       Age        SibSp   Parch      Ticket               Fare        Embarked
##  Min.   : 0.42   0:608   0:678   Length:891         Min.   :  0.00   C:170
##  1st Qu.:22.00   1:209   1:118   Class :character   1st Qu.:  7.91   Q: 77
##  Median :29.70   2: 28   2: 80   Mode  :character   Median : 14.45   S:644
##  Mean   :29.70   3: 16   3:  5                      Mean   : 32.20
##  3rd Qu.:35.00   4: 18   4:  4                      3rd Qu.: 31.00
##  Max.   :80.00   5:  5   5:  5                      Max.   :512.33
##                  8:  7   6:  1
##         Title         Fsize     FsizeD
##  Master     : 40   1      :537   1:537
##  Miss       :185   2      :161   2:292
##  Mr         :517   3      :102   3: 49
##  Mrs        :126   4      : 29   4: 13
##  Rare Title: 23    6      : 22
##                    5      : 15
##                    (Other): 25
```

```
#Características de las variables
str(titanic_data)
```

```
## 'data.frame':    891 obs. of  14 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
```

```
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : Factor w/ 7 levels "0","1","2","3",..: 2 2 1 2 1 1 1 4 1 2 ...
##  $ Parch      : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1 2 3 1 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ Title      : Factor w/ 5 levels "Master","Miss",..: 3 4 2 4 3 3 3 1 4 4 ...
##  $ Fsize      : Factor w/ 9 levels "1","2","3","4",..: 2 2 1 2 1 1 1 5 3 2 ...
##  $ FsizeD     : Factor w/ 4 levels "1","2","3","4": 2 2 1 2 1 1 1 3 2 2 ...
```

Análisis de los estadísticos descriptivos:

- Al abrir la tabla de datos sin procesar, se puede constatar que hay valores faltantes en las columnas 'Age' (177), y 'Embarked' (2).
- Las siguientes variables tienen valores discretos: 'Survived', 'Pclass', 'Sex', 'Embarked', 'SibSp', 'Parch' y 'Title', por lo cual fueron convertidas a factores para poder visualizar la frecuencia absoluta de cada uno de sus niveles.
- La variables 'Age' y 'Fare' son numéricas de carácter continuo.
- Las variables 'Name' y 'Ticket' son del tipo carácter, por lo cual no se pueden obtener estadísticos descriptivos de ellas. Bajo esta premisa, son prescindibles.

Propuesta para discretizar el tamaño de la familia:

- Considero discretizar el tamaño de la familia en los siguientes grupos:
  - Familias de 1 pasajero: que representan el 60% del total de pasajeros.
  - Familias de 2 a 4 pasajeros: debido a que son quienes presentan mayores opciones de permanecer con vida.
  - Familias de 5 a 7 pasajeros: los cuales tienen un porcentaje de supervivencia muy parecido entre sí, pero inferior al grupo anterior.
  - Familias de más de 8 o más pasajeros: quienes tienen nula opción de supervivencia (ver segunda tabla en página 10).
- El resultado de la discretización muestra que quienes tienen más opciones de sobrevivir son los grupos familiares de 2 a 4 miembros (con un 58%), seguidos de pasajeros que viajaban solos (con un 30%). En penúltima posición se encuentra el grupo de familias de 5 a 7 miembros (con un 20% de probabilidad de sobrevivir), para finalizar con los grupos de 8 o más integrantes (con 0%).

Mecanismos para la generación de datos faltantes

- Para la variable 'Embarked', otra forma de lidiar con datos faltantes sería omitirlos (**), ya que tienen un peso del 0,22% respecto al total (2 de 891 observaciones). O bien, se les podría asignar aleatoriamente alguna de las 3 categorías ("C", "Q" o "S").
- Para la variable 'Age', aparte de la solución propuesta (completar datos faltantes reemplazándolos con la media de la edad), podría evaluarse el siguiente caso:

- Generar 'x' números aleatorios (en este caso 177, por la cantidad de valores 'NA'), siguiendo una distribución normal.
- Con este método, el resultado conserva la distribución original de los datos en la columna 'Age'.

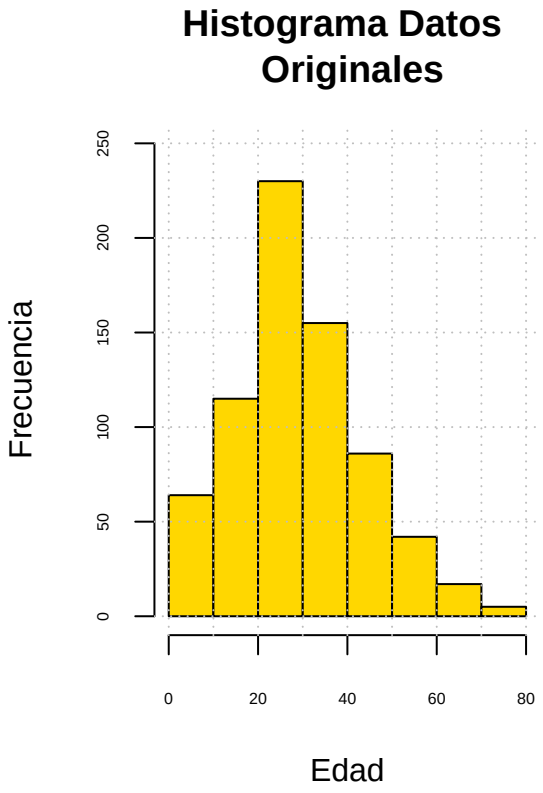
```
#Se crea una nueva tabla para no sobre-escribir la anterior
titanic2 <- read.csv("../Archivos R/titanic.csv",header=TRUE)
#Se calcula la media de los datos originales
mean2 <- round(mean(na.omit(titanic2$Age)),2)
#Se calcula la desviación estándar
desv2 <- round(sd(na.omit(titanic2$Age)),2)

#Se generan números aleatorios que sigan una distribución normal,
#los cuales reemplazarán a los valores NA de la columna 'Age'
random_NA<-abs(rnorm(n=177,mean=mean2,sd=desv2))

#Se crea una variable para guardar la columna 'Age'
Age_random_NA <- data.frame(titanic2$Age)
#Se asigna a los valores NA los datos guardados de la variable que
#generó los números aleatorios
Age_random_NA[is.na(Age_random_NA)] <- random_NA

#Comparación de estadísticos: datos originales
summary(na.omit((titanic2$Age)))
```

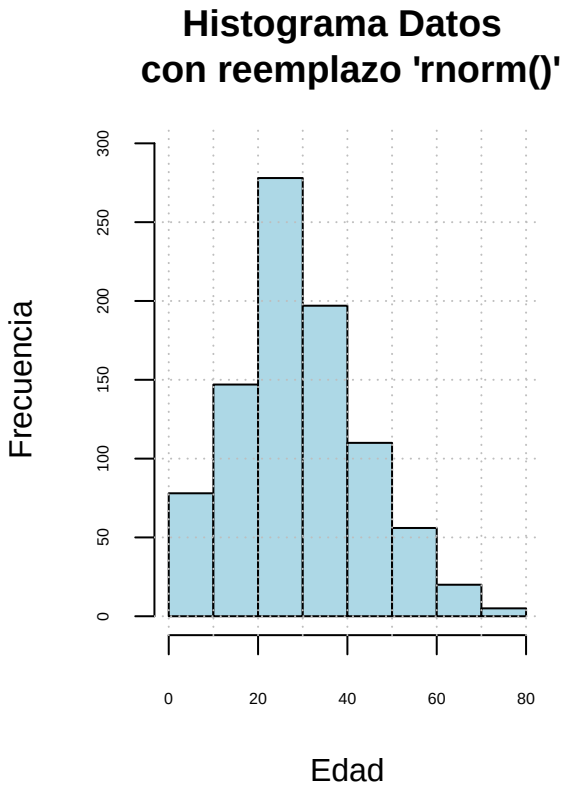
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   20.12   28.00   29.70   38.00   80.00
```

```
#Comparación de estadísticos: datos con reemplazo de valores 'NA'
summary(Age_random_NA$titanic2.Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   20.00   28.00   29.68   38.33   80.00
```

```
#Gráficos de histograma mostrados en una fila y dos columnas
par(mfrow=c(1,2))
#Histograma de los datos originales
hist(na.omit((titanic2$Age)),
main = "Histograma Datos \n Originales",
ylab = "Frecuencia",
xlab = "Edad",
ylim = c(0, 250),
col="gold",
cex.axis=0.5)
abline(h=seq(0,250,50), v = seq(0,80,10), lty="dotted", col = "grey")
#Hisotrama de los datos con reemplazo random
hist(Age_random_NA$titanic2.Age,
main = "Histograma Datos \n con reemplazo 'rnorm()'",
ylab = "Frecuencia",
xlab = "Edad",
ylim = c(0, 300),
col = "light blue",
cex.axis=0.5)
abline(h=seq(0,250,50), v = seq(0,80,10), lty="dotted", col = "grey")
```






```
#La nueva representación gráfica conserva la distribución de los datos originales
```

(**) [Parte 1: Preprocesamiento de Datos, Datos faltantes, Diapositiva 8]