Pontificia Universidad Católica de Chile

FORO Nº 1
Introducción al Modelamiento Estadístico y Sistemas Recomendadores

"Procesamiento de Datos"

Eduardo Andrés Carrasco Vidal
Ingeniero Civil Industrial

Actividad Nº 1:

En esta actividad usted deberá aportar con 1 respuesta principal a la o las preguntas enunciadas por el profesor y comentar y/o intervenir 2 respuestas de sus compañeros (en otros hilos de conversación). Además, junto a la respuesta principal, usted deberá subir un informe del trabajo realizado, el cual también será considerado en la evaluación del foro.

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

    a) Cargue el conjunto de datos en la sesión de trabajo R, usando la función *read.table()*:

```
> ?read.table
> datos <- read.table(file.choose(),header=TRUE, sep=",") #cargar datos
> View(datos)
> summary(datos) #estadística descriptiva
 hour_per_week
 Min.   : 1.00
 1st Qu.:40.00
 Median :40.00
 Mean   :40.44
 3rd Qu.:45.00
 Max.   :99.00
>install.packages("Hmisc")
> library("Hmisc")
> describe(datos) #estadística descriptiva con librería "Hmisc" con detalle de observación#
datos

 1  Variables      32561  Observations
---------------------------------------------------------------------------------------------------------------
hour_per_week
      n   missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90
 32561         0       94    0.897    40.44    12.28       18       24       40       40       45       55
     .95
      60

lowest :  1  2  3  4  5, highest: 95 96 97 98 99
---------------------------------------------------------------------------------------------------------------
```

b) Calcule en forma manual el puntaje Z, para las horas de trabajo semanal:

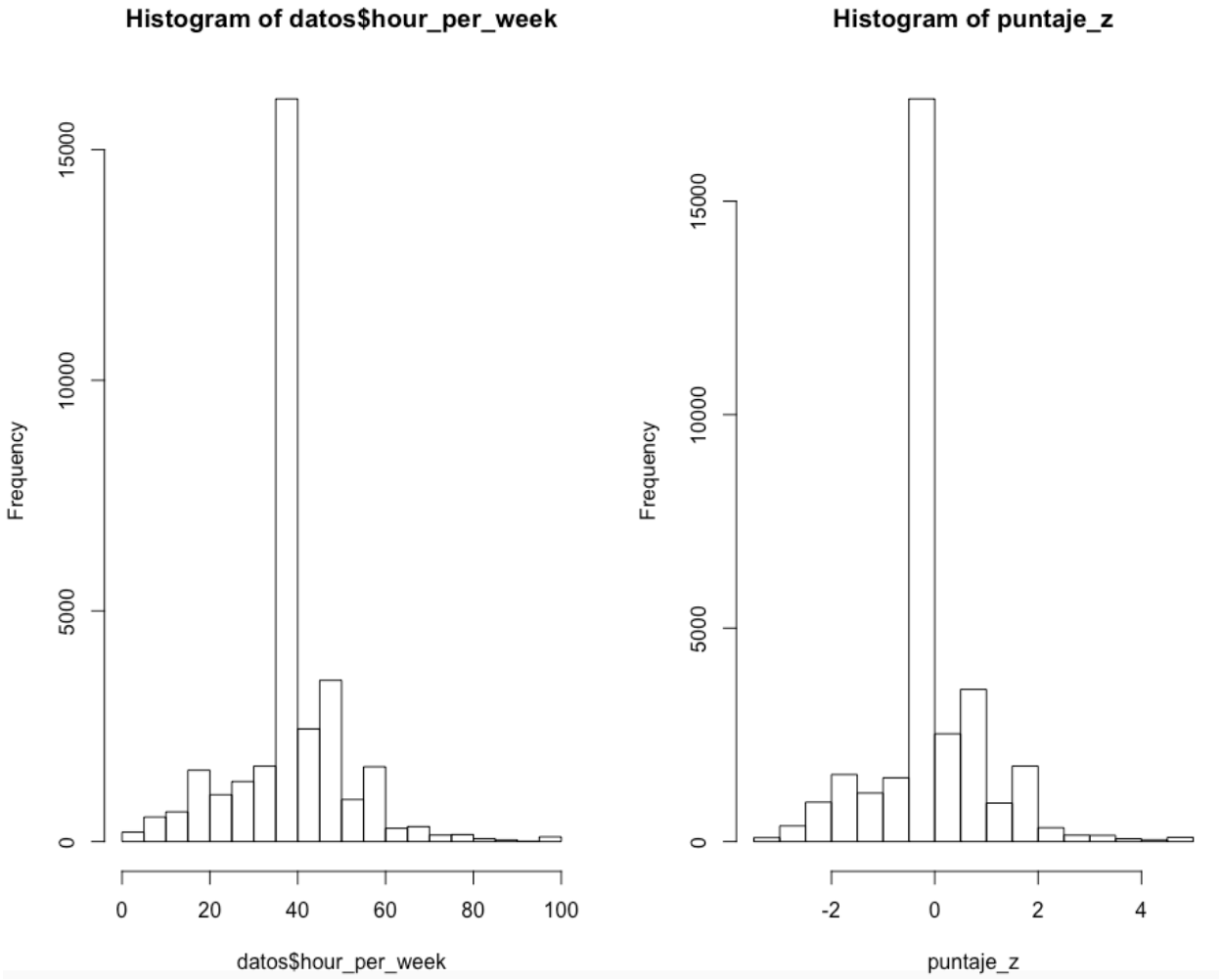
```
> media.o <- mean(datos$hour_per_week) #calcular la media aritmética
> desv.o <- sd(datos$hour_per_week) #calculo de la desviación estándar
> media.o #mostrar la media
[1] 40.43746
> desv.o #mostrar la desviación estándar
[1] 12.34743
> puntaje_z <- (datos$hour_per_week-media.o)/desv.o #calculo manual puntaje z
> summary(puntaje_z) #estadística descriptiva puntaje z
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.19398 -0.03543 -0.03543  0.00000  0.36951  4.74289
> describe(puntaje_z) #estadística descriptiva puntaje z, usando libreria "Hmisc"
puntaje_z
       n    missing   distinct       Info       Mean        Gmd        .05        .10        .25        .50
   32561          0         94      0.897 -3.995e-17     0.9947   -1.81718   -1.33125   -0.03543   -
0.03543
      .75        .90        .95
  0.36951    1.17940    1.58434

lowest : -3.193981 -3.112993 -3.032004 -2.951016 -2.870027, highest:  4.418940
4.499928  4.580917  4.661905  4.742894
> media.z <- mean(puntaje_z) #calculo media puntaje z
> desv.z <- sd(puntaje_z) #calculo desviación estandar puntaje z
> media.z #mostrar media
[1] -3.883133e-17
> desv.z #mostrar desviación estándar
[1] 1
```

c) 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:

```
> par(mfrow=c(1,2))
> hist(datos$hour_per_week)
> hist(puntaje_z)
```

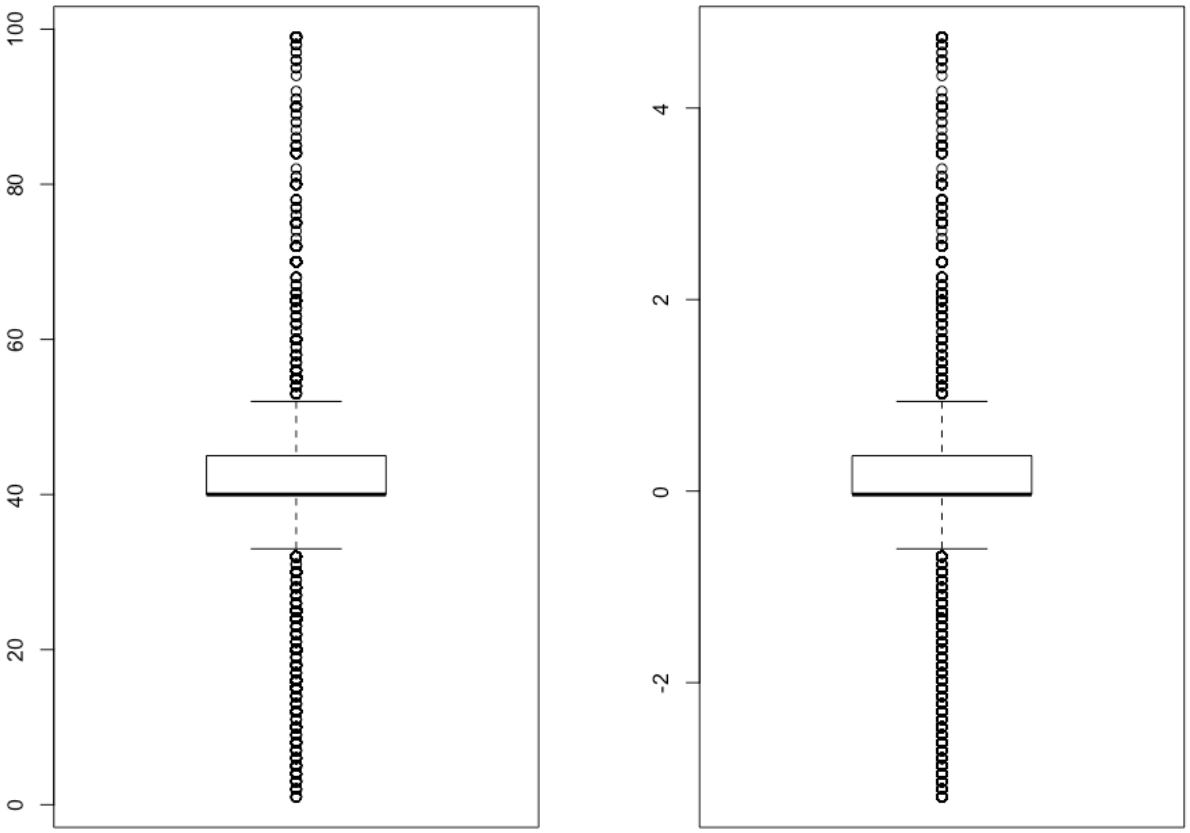
De acuerdo a lo observado en ambos histogramas, podemos mencionar respecto a los **datos originales** que en el espacio muestral n = 32.561, con un rango de [min 1: max 99], existe una media aritmética de 40.44, ligeramente superior a la mediana de 40, mostrando una asimetría positiva con alta presencia de números bajos en la muestra, ubicando una moda de alta frecuencia con más de 15000 eventos cercanos a las 40, considerando además la desviación estándar de 12,35, lo cual, podría señalar que existe una tendencia mayoritariamente Uni-modal, a pesar que el histograma presenta más de un pico en el gráfico.
Respecto a los **datos estandarizados**, cuyo espacio muestral es el mismo (n = 32.561), posee un rango de [min -3,20 : max 4,75], existe una media aritmética de -3,89e-17 inferior a la mediana de -0,036 (consideradas casi 0) mostrando de igual

manera una asimetría positiva con presencia de números bajos en el histograma; si se observa la frecuencia de eventos y se analiza la desviación estándar = 1, se puede señalar que al igual que el set de datos originales, existe una tendencia mayoritariamente Uni-modal, pero este histograma presenta menor cantidad de picos en el gráfico debido a que el puntaje z permite expresar los valores en unidades de desviación estándar analizando tendencias centrales.



d) 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:

```
> par(mfrow=c(1,2))
> boxplot(datos$hour_per_week)
> boxplot(puntaje_z)
```

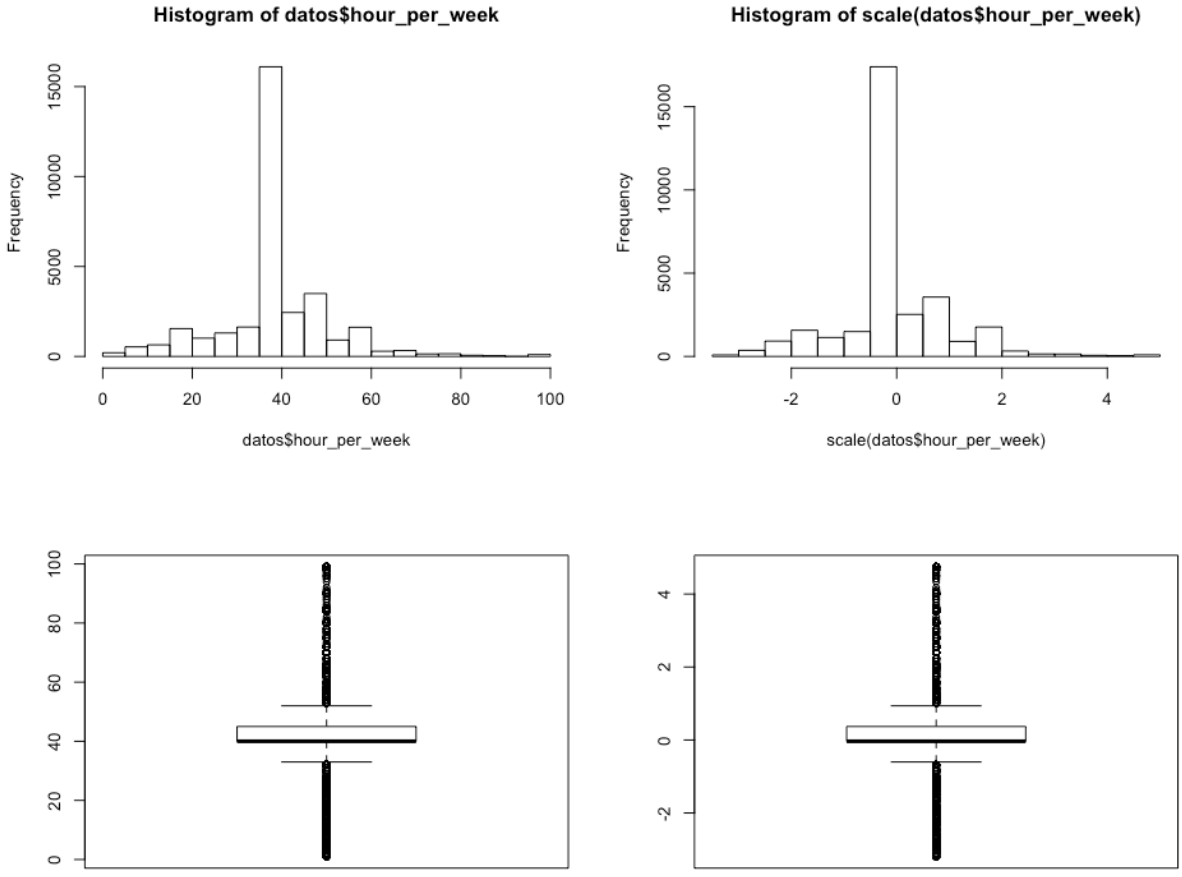


Si observamos los gráficos boxplot y los resultados obtenidos previamente con la función *summary()* para los **valores originales**, tenemos en el primer cuartil un valor de 40 mientras que en el tercer cuartil un valor de 45, con lo cual se obtiene un IQR (Inter-Quartil Range) equivalente a 5, que es la medida de la caja central (cuadrilatero). También se pueden distinguir los límites tanto inferiores como superiores del boxplot permitiendo identificar los valores atípicos que superan el rango entre [32,5 : 47,5], considerando un valor de aproximadamente 1.5 veces el IQR; de lo anterior, todos los valores que se encuentran fuera de este rango, se consideran valores *"outliers"*.

Si analizamos el gráfico boxplot de los **datos estandarizados** (puntaje z), utilizando la misma función, tenemos en el primer cuartil un valor de -0,035 mientras que en el tercer cuartil un valor de 0,37, con lo cual se obtiene un IQR equivalente a 0,405 que es la medida de la caja central. También se pueden distinguir los límites tanto inferiores como superiores del boxplot permitiendo identificar a todos los valores que se encuentren fuera de este rango como "*outliers*", considerando a estos valores como medida de la desviación estándar.

e) Repita los pasos c) y d), usando la función *scale()*:

```
> par(mfrow=c(2,2))
> hist(datos$hour_per_week)
> hist(scale(datos$hour_per_week))
> boxplot(datos$hour_per_week)
> boxplot(scale(datos))
```



Como se observa, la función *scale()* permite estandarizar los valores, obteniendo los mismos resultados que la aplicación del puntaje z de manera manual, como se observa en el siguiente cuadro:

```
> est<-scale(datos$hour_per_week)
> summary(est)
       V1
 Min.   :-3.19398
 1st Qu.:-0.03543
 Median :-0.03543
 Mean   : 0.00000
 3rd Qu.: 0.36951
 Max.   : 4.74289
> summary(puntaje_z)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.19398 -0.03543 -0.03543  0.00000  0.36951  4.74289
```

2. Considere los datos 'titanic.csv', sobre la tragedia del Titanic, cuya descripción se muestra en la siguente 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)

a) Cargue el conjunto de datos en la sesión de trabajo R, usando la función *read.table()*:

```
> ?read.table
> datos <- read.table(file.choose(),header=TRUE, sep=",") #cargar datos
> View(datos)
```

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

```
> summary(datos)
  PassengerId       Survived          Pclass                                      Name         Sex
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                  :  1   female:314
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward          :  1   male  :577
 Median :446.0   Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)     :  1
 Mean   :446.0   Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                  :  1
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):  1
 Max.   :891.0   Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin       :  1
                                                  (Other)                              :885
      Age            SibSp           Parch             Ticket         Fare        Embarked          Title
 Min.   : 0.42   Min.   :0.000   Min.   :0.0000   1601    :  7   Min.   :  0.00   C   :168   Master     : 40
 1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   347082  :  7   1st Qu.:  7.91   Q   : 77   Miss       :185
 Median :28.00   Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45   S   :644   Mr         :517
 Mean   :29.70   Mean   :0.523   Mean   :0.3816   3101295 :  6   Mean   : 32.20   NA's:  2   Mrs        :126
 3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000   347088  :  6   3rd Qu.: 31.00              Rare Title: 23
 Max.   :80.00   Max.   :8.000   Max.   :6.0000   CA 2144 :  6   Max.   :512.33
 NA's   :177                                      (Other) :852
```

De acuerdo a los resultados obtenidos en esta función, podemos verificar la cantidad total de pasajeros registrados (**PassengerId**) equivalente a 891 personas, de los **Survived**, al utilizar una variable lógica (1=sobreviviente, 0=no sobreviviente) la función summary no es util, debiendo utilizar una tabla de frecuencias para determinar la cantidad de personas sobrevivientes; de igual manera con la variable

Pclass (Clase del pasajero), en la cual existen 3 tipos correspondientes a una frecuencia (1ra, 2da y 3ra clase). Se puede determinar además respecto a las variables cualitativas (**Name** y **Sex**) que no se registran nombres que se repitan y que además, del listado de pasajeros, tenemos 314 mujeres y 577 hombres.
Respecto a la edad (**age**), el menor es de 0.42 años hasta el mayor de 80 años, considerando que existen 177 pasajeros cuya edad no fue informada.
De la variable **Sibsp** (Número de hermanos o cónyuges a bordo), podemos observar que los pasajeros tenían hermanos/cónyuges a bordo con un mínimo de 0 y un máximo de 8 parientes de este tipo; también la variable **Parch** (Número de padres o hermanos a bordo), permite observar los pasajeros que tenían padres/hijos a bordo con un mínimo de 0 y un máximo de 6 parientes de este tipo.
La variable **Ticket**, representa el boleto con el cual ingresó cada pasajero y se divide en una tabla de frecuencias determinada para cada valor, el precio (**fare**) pagado por cada ticket tiene un rango desde los $0 a $512.33.
La variable **Embarked** representa el puerto de embarque (Embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton), donde 168 entraron por el muelle C, 77 entraron por el muelle Q, 644 por el muelle S y 2 pasajeros de los cuales no se tiene información del muelle de subida.

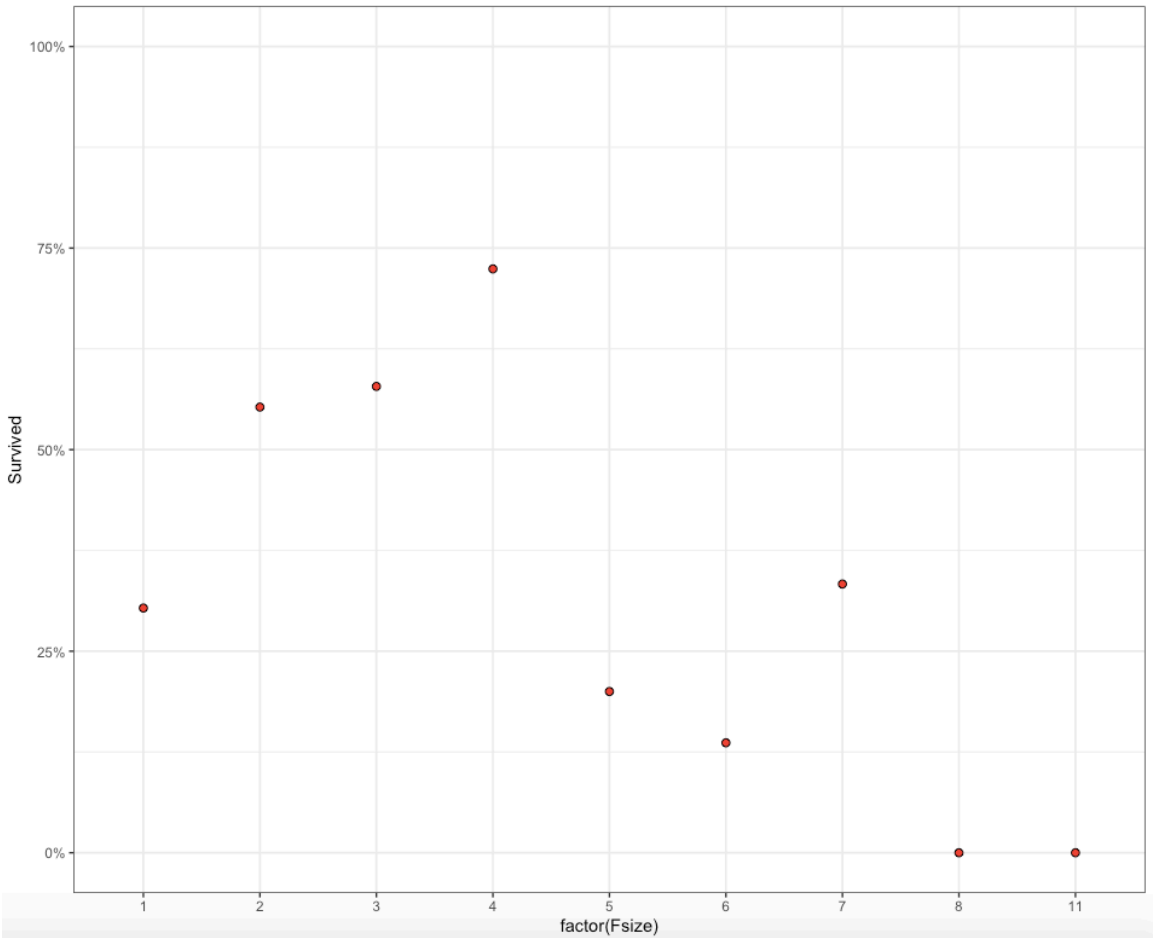
c) Cree una variable que indique el tamaño total de la familia del pasajero (incluyéndose el mismo):

```
datos$Fsize <- datos$SibSp + datos$Parch + 1
```

Esta función permite crear una nueva variable en el conjunto principal datos que suma la cantidad de hermanos y esposas, más la cantidad de padres e hijos, además de sumarse a sí mismo.

d) Grafique la relación entre la tasa de sobrevivientes y el tamaño de la familia:

```
> ggplot(datos, aes(factor(Fsize), Survived)) +
+     stat_summary(fun.y=mean, geom="point", shape=21, fill="red", size=2) +
+     scale_y_continuous(labels=percent_format(), limits=c(0,1)) +
+     theme_bw()
```



De acuerdo al gráfico, no hay una relación directa entre la cantidad de parientes (miembros de la familia) y el porcentaje de sobreviviencia. Sólo se puede señalar que aquellos que tienen 4 parientes, tienen una tasa de sobrevivencia cercana al 75% y así con el resto de los factores.

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

Si consideramos el proceso de dicretización como una técnica de transformación de datos que permita abordar el proceso de minería de forma más eficiente, podemos establecer categorías para la clasificación de variables numéricas.
Por lo cual, el proceso de clasificación dependerá sólo del usuario o de la forma en que le parezca mejor trabajar, cumpliendo con la lógica del problema y los resutados esperables. Es por esto, que se proponen las siguientes clasificaciones:

Solitario: 1 persona.
Familia mínima: 2 a 4 personas (padre, madre, hijos hasta 2).
Familia mediana: 5 a 7 personas (padre, madre, hijos, hermanos).
Familia grande: 7 personas o más (padre, madre, hijos, hermanos).

```
> datos$FsizeD[datos$Fsize == 1] <- "Solitario"
> datos$FsizeD[datos$Fsize == 1] <- "Solitario"
> datos$FsizeD[datos$Fsize < 5 & datos$Fsize > 1] <- "Familia_Minima"
> datos$FsizeD[datos$Fsize < 8 & datos$Fsize > 4] <- "Familia_Mediana"
> datos$FsizeD[datos$Fsize > 7] <- "Familia_Grande"
> table(datos$FsizeD)

 Familia_Grande Familia_Mediana  Familia_Minima       Solitario
             13              49             292             537
```

f) 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**:

```
> Embarked_NA <- datos[is.na(datos$Embarked),]
> Embarked_NA$PassengerId
[1]  62 830 #representa que los pasajeros Nº 62 y 830, no tienen ingresada su puerta de
embarque
> Age_NA <- datos[is.na(datos$Age),]
> summary(Age_NA$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
     NA      NA      NA     NaN      NA      NA     177
> summary(Age_NA$Sex)
female   male
    53    124
```

Como se observa en el recuadro anterior, respecto del puerto de embarque (**Embarked**) existen 2 pasajeros (Nº 62 y Nº 830) de los cuales no se tiene información, de igual manera, existen 177 pasajeros que no poseen un registro de edad.

En este caso de datos faltantes, la falta de datos es independiente de los datos observados y no observados, ocurriendo completamente al azar (puede que no se haya podido registrar o se perdió la información en el hundimiento), cualquier información tiene igual probabilidad de perderse, por lo cual, corresponde al tipo *"Missing Completely at Random"* (MCAR), debiendo utilizar el mecanismo de imputación de datos faltantes.

Para el caso de la variable **embarked**, los datos faltantes representan menos del 0,21% por lo cual es válido asignar una variable en forma aleatoria, no afectando los valores finales; por otra parte, en el caso de la variable **age**, sustituir las observaciones perdídas con la media no añade ninguna información nueva, la media global es la misma y se subestima el error. En resumen, ambos procedimientos son válidos para este tipo de datos faltantes.

g) Genere un conjunto de datos completos al imputar los valores faltantes de la variable *embarked* por las de la puerta "C" y los valores faltantes de la variable *age*, por el promedio de edad de los datos observados:

```
> datos$Embarked[Embarked_NA$PassengerId] <- 'C'
> summary(datos$Embarked)
  C   Q   S
170  77 644
>datos$Age[Age_NA$PassengerId] <- mean(na.omit(datos$Age))
> summary(datos$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.42   22.00   29.70   29.70   35.00   80.00
```

La ejecución de los comandos anteriores, nos permiten verificar que la asignación de valores a los Na (sin valor), está imputada y fue reemplazada por valores de "C" y del promedio de edad (mean) de las edades observadas.

REFERENCIAS:

1. Shankay45 (2016) Diagrama de Cajas. Sitio Picando con R, R en español. [en línea] Recuperado de: https://picandoconr.wordpress.com/2016/02/27/diagrama-de-cajas/
2. Universidad de Alicante (s.f.) Características de las distribuciones de frecuencia. Sitio Repositorio Institucional, Universidad de Alicante. [en línea] Recuperado de: https://rua.ua.es/dspace/bitstream/10045/18578/1/Capitulo%203.pdf
3. Alcaraz, Francisco (2013) Clasificación y ordenación con R. Sitio: Universidad de Murcia, España. [en línea] Recuperado de: https://www.um.es/docencia/geobotanica/ficheros/practica2.pdf
4. Mandeville, Peter (2010) Observaciones Perdidas. Sitio: Fundación Dialnet. [en línea] Recuperado de: https://dialnet.unirioja.es/servlet/articulo?codigo=3245988