Complementos

Sesión 8

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

Customización de gráficos en R

Visualización de la información

Los gráficos muchas veces nos ayudan a observar y comprender determinado fenómeno. Cuando un gráfico está bien realizado, es claro, legible, interpretable e informativo.

Datos hotel

La base de datos hotel contiene información sobre todas las reservaciones de un hotel:

- **Cancel**: Indica si la reserva fue cancelada o no (1: Cancelada, 0: No cancelada)
- **Mes**: Mes de la reserva
- **Weekend**: Cantidad de días de fin de semana (Sábados o Domingos) reservados
- **Weekday**: Cantidad de días de la semana (Lunes, Martes, Miércoles, Jueves, Viernes) reservados

Práctica 1

a) Cargue los datos. Recodifique la variable Cancel como corresponda. ¿Cuál es el porcentaje de reservas canceladas?

b) ¿En qué meses se realizaron más cancelaciones? Muestre la información en un gráfico de torta. Para realizar un gráfico de torta, utilice:

```
pie(table(hotel$Mes[which(hotel$Cancel=="Cancelada")]))
```

Explore los argumentos que puede modificar en esta función corriendo `?pie`. Modifique los colores del gráfico y añada el título.

Práctica 1

c) Instale y cargue el paquete `plotrix`. Y realice el siguiente gráfico de torta 3D:

```
frecuencias<-data.frame(table(hotel$Mes[which(hotel$Cancel=="Cancelada")]))
slices <- frecuencias$Freq
lbls <-frecuencias$Var1
pie3D(slices,labels=lbls, Función pie3D del paquete plotrix crea el gráfico main="Grafico de
torta de cancelaciones por mes", col=rainbow(12, alpha=0.4), theta=pi/3, labelcex=0.8)
```

Pruebe distintos valores para los argumentos theta y labelcex. ¿Qué especifican estos argumentos

Práctica 1

Práctica 2

El NEIC (*National Earthquake Information Center*) determina la ubicación y el tamaño de los sismos importantes que ocurren en todo el mundo. La base de datos `terremotos` contiene registros de sismos desde el año 1999 hasta el año 2016. La información contenida en la base de datos es la siguiente:

- `ID` Identificador del sismo
- `Date` Fecha del sismo
- `Time` Hora del sismo
- `Latitude` y `Longitud` indican ubicación del el epicentro del sismo
- `Type` Tipo del sismo
- `Depth` Profundidad del sismo
- `Magnitude` Magnitud del sismo
- `Magnitude Type` Escala utilizada para medir la magnitud del sismo

Práctica 2

a) ¿Cómo se distribuye la cantidad de sismos por año? ¿Se ha visto un aumento en la cantidad de sismos en el tiempo? Muestre la información en un gráfico.

b) ¿Se observan diferencias en la magnitud del sismo dependiendo de la escala de medición? Muestre la información en un gráfico, comente.

Respuestas práctica 1

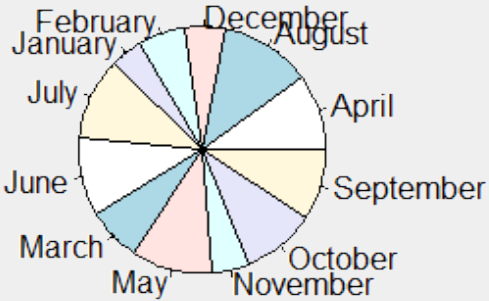
Respuestas práctica 1 a)

```
library(rio) #Carga el paquete rio

#Defino la carpeta donde se encuentra el archivo (el directorio)
setwd("C:/Users/HP/Desktop/Trabajo/2020-2/Laboratorio I. Estadística/Sesiones-Semanas/S9 Variables continuas")

hotel<-import('hotel.xlsx') #Cargo los datos hotel.xlsx que están en la carpeta

str(hotel)
'data.frame': 118987 obs. of  4 variables:
 $ Cancel : num  0 0 0 0 0 0 0 0 1 1 ...
 $ Mes    : chr  "July" "July" "July" "July" ...
 $ Weekend: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Weekday: num  0 0 1 1 2 2 2 2 3 3 ...

hotel$Cancel<-ifelse(hotel$Cancel==0, "No cancelada", "Cancelada") #Recodifica en categorías

table(hotel$Cancel) #Cantidad de reservas canceladas y no canceladas
   Cancelada No cancelada
       44115        74872

table(hotel$Cancel)/nrow(hotel) #Proporción de reservas canceladas y no canceladas
   Cancelada No cancelada
   0.3707548    0.6292452

table(hotel$Cancel)/nrow(hotel)*100 #Porcentaje
   Cancelada No cancelada
    37.07548     62.92452

#El 37% de las reservaciones fue cancelada. Es un valor bastante alto!
#Lo ideal es que las reservas no sean canceladas.
```

Respuestas práctica 1 b)

```
table(hotel$Mes[which(hotel$Cancel=="Cancelada")]) #Muestra cuántas cancelaciones hay por mes

    April    August  December  February   January      July      June
     4510      5226      2359      2687      1802      4727      4531
    March       May  November   October September
     3137      4669      2115      4244      4108

tabla<-data.frame(table(hotel$Mes[which(hotel$Cancel=="Cancelada")])) #Guardo la info en una tabla

head(tabla[order(tabla$Freq, decreasing=TRUE),],5) #5 Meses con más cancelaciones
    Var1 Freq
2 August 5226
6   July 4727
9    May 4669
7   June 4531
1  April 4510

#El mes con mas cancelaciones fue Agosto con 5226 reservas canceladas.
#Luego le sigue Julio con 4727 y así respectivamente con Mayo, Junio y Julio.

pie(table(hotel$Mes[which(hotel$Cancel=="Cancelada")])) #Gráfico con la info
```

Respuestas práctica 1 b)



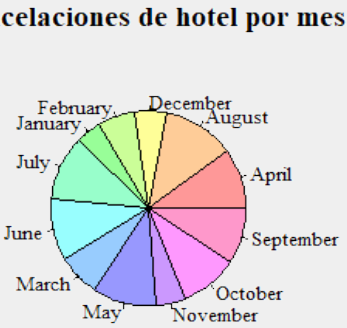
Respuestas práctica 1 b)

```
#col= especifica los colores
#main= especifica el titulo

#rainbow(12, alpha=0.4) genera un vector con 12 colores, alpha indica la transparencia
pie(table(hotel$Mes[which(hotel$Cancel=="Cancelada")]), col=rainbow(12, alpha=0.4), main="Cancelaciones de hotel por mes")
```

Respuestas práctica 1 b)

Respuestas práctica 1 b)

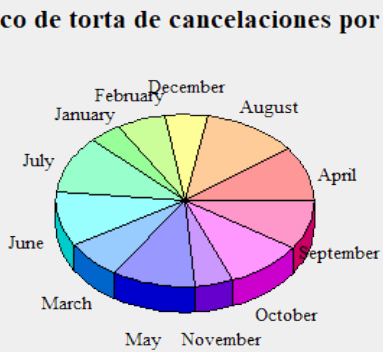
```
#Cambio el tipo de letra:

library(extrafont) #Carga el paquete extrafont
fonts() #Muestra los tipos de letra disponibles

par(family="Times New Roman")

#Con cex=0.8 se modifica el tamaño de los labels (etiquetas) mas pequeño
pie(table(hotel$Mes[which(hotel$Cancel=="Cancelada")]), col=rainbow(12, alpha=0.4), main="Cancelaciones de hotel por mes", cex=0.8)
```



Respuestas práctica 1 c)

```
#install.packages("plotrix") #Instala el paquete plotrix
library(plotrix) #Carga el paquete plotrix

frecuencias<-data.frame(table(hotel$Mes[which(hotel$Cancel=="Cancelada")])) #Guarda info en una tabla

slices <- frecuencias$Freq #Cantidad de cancelaciones por mes
lbls <-frecuencias$Var1 #Nombre de los meses

pie3D(slices,labels=lbls, #Función pie3D del paquete plotrix crea el gráfico
      main="Grafico de torta de cancelaciones por mes", col=rainbow(12, alpha=0.4), theta=pi/3, labelcex=0.8)

#theta modifica el angulo
#labelcex el tamaño de los labels igual que cex en la función pie()
```

Respuestas práctica 1 c)

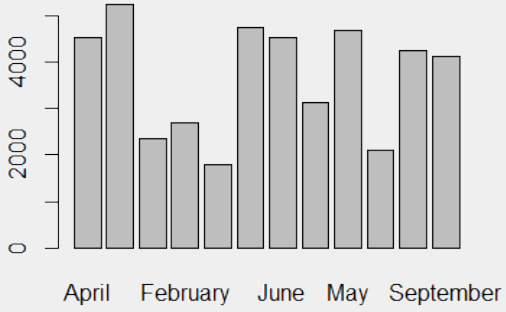


Nota: Hay que tener mucho cuidado con los gráficos 3D, de modo que no pierdan interpretación o no vayan a generar falsas sensaciones visuales de la información.

Respuestas práctica 1 c)

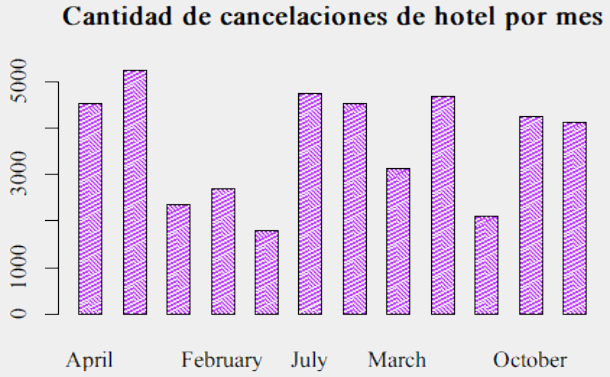
La información también se podía haber mostrado en un gráfico de barras:

```
barplot(table(hotel$Mes[which(hotel$Cancel=="Cancelada")]))
```



Respuesta práctica 1 c)

```
par(family="Mongolian Baiti") #Tipo de letra

barplot(table(hotel$Mes[which(hotel$Cancel=="Cancelada")]),
        col="darkorchid1",#color
        space=0.9, #Espacio entre barras
        density=60, #Diseño coloreado de las barras
        main="Cantidad de cancelaciones de hotel por mes" #Titulo
        )
    #horiz=TRUE es para ver las barras de forma horizontal
```

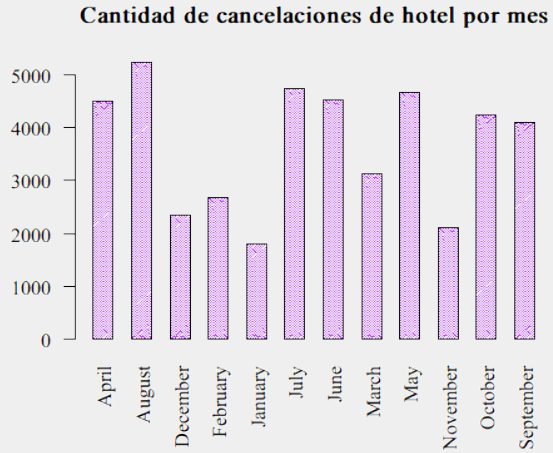


Editando labels en un gráfico de barras

En el gráfico anterior no se observan bien los nombres de los meses :(podemos editar la orientación de estos nombres de modo que aparezcan:

```
par(family="Mongolian Baiti") #Tipo de letra

barplot(table(hotel$Mes[which(hotel$Cancel=="Cancelada")]),
        col="darkorchid1",#color
        space=0.9, #Espacio entre barras
        density=60, #Diseño coloreado de las barras
        main="Cantidad de cancelaciones de hotel por mes", #Titulo
        las=2 #Cambia orientación
)
```

Editando labels en un gráfico de barras

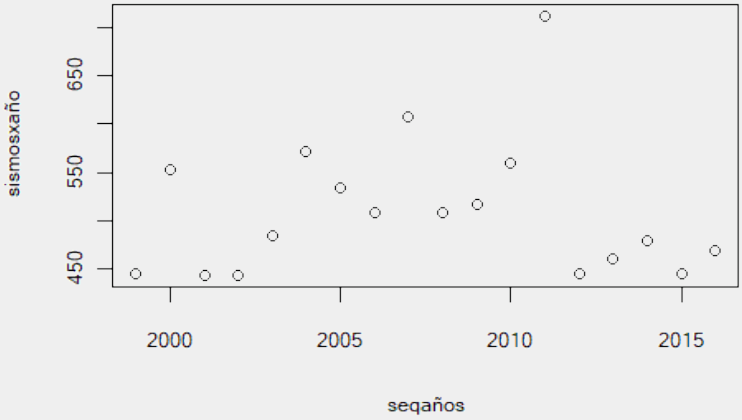


Respuestas práctica 2

Respuestas práctica 2 a)

```
sismos<-import('sismos.txt')

str(sismos$Date) #Hay que extraer el año
chr [1:9189] "01-02-1999" "01-04-1999" "01-05-1999" ...

sismos$Year<-as.numeric(substr(sismos$Date, start=7, stop=10))
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011
 446  553  443  444  485  571  533  508  608  508  517  560  712
2012 2013 2014 2015 2016
 445  461  480  446  469

as.vector(table(sismos$Year)) #Vectoriza la info
 [1] 446 553 443 444 485 571 533 508 608 508 517 560 712 445 461
[16] 480 446 469

sismosxaño<-as.vector(table(sismos$Year))
seqaños<-1999:2016

par(family="Segoe MDL2 Assets")
plot(seqaños, sismosxaño)
```

Respuestas práctica 2 a)

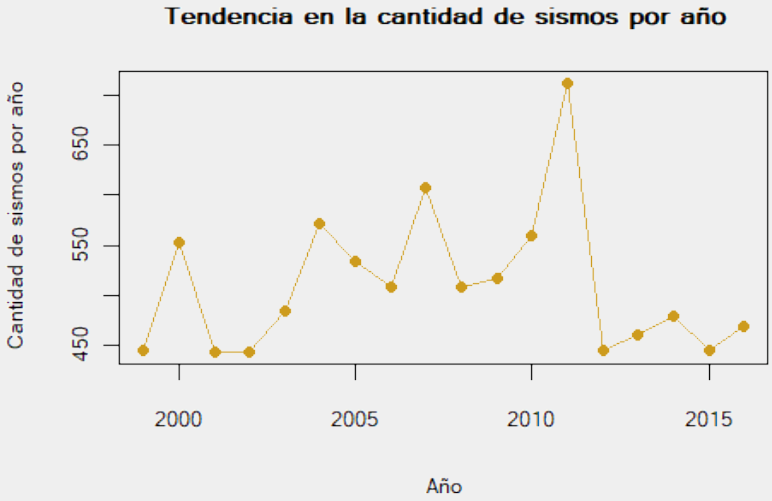


 no son adecuados, falta título y

Respuestas práctica 2 a)

```
par(family="Segoe MDL2 Assets")
plot(seqaños, sismosxaño,
     type="o", #Gráfico de línea con puntos, una "l" sería sólo línea
     xlab="Año", #Nombre eje x
     ylab="Cantidad de sismos por año", #Nombre eje y
     main="Tendencia en la cantidad de sismos por año", #Titulo
     pch=19, #Rellena los circulos de negro
     col="goldenrod3" #Color
     )
```

Respuestas práctica 2 a)



Respuestas práctica 2 b)

Resulta natural esperar diferencias en la magnitud del sismo si éste se pide en base a una escala u otra, pues las distintas escalas para medir sismos se basan en distintas cosas, percepción, daños a edificios, etcétera. Pero resulta interesante qué tanto varían las mediciones por escala de sismo.

Respuestas práctica 2 b)

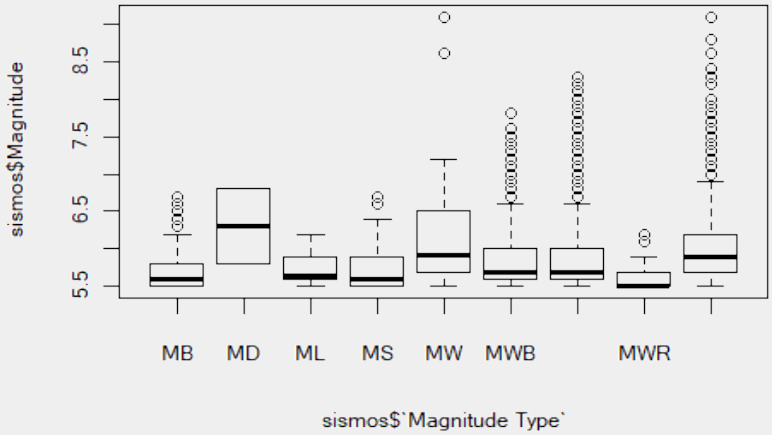
Primero, hay que determinar cuántas escalas de sismos hay en los datos:

```
unique(sismos$‘Magnitude Type‘) #Muestra todas las escalas en los datos
[1] "MWC" "MWB" "MB"  "MW"  "MD"  "ML"  "MS"  "MWW" "MWR"

length(unique(sismos$‘Magnitude Type‘)) #Hay 9 escalas
[1] 9
```

Podemos hacer un gráfico para cada escala:

```
par(family="Segoe MDL2 Assets")
boxplot(sismos$Magnitude~sismos$‘Magnitude Type‘)
```

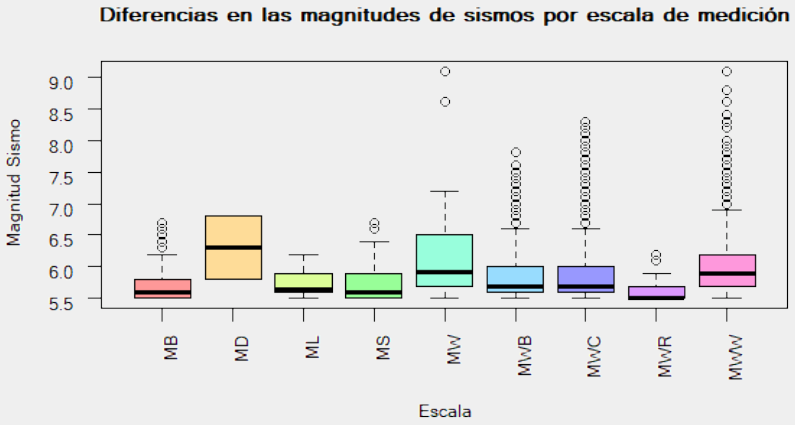
Respuestas práctica 2 b)



Respuestas práctica 2 b)

```
par(family="Segoe MDL2 Assets")
boxplot(sismos$Magnitude~sismos$`Magnitude Type`,
        col=rainbow(9, alpha=0.4), #Colores
        xlab="Escala", #Nombre eje x
        ylab="Magnitud Sismo", #Nombre eje y
        main="Diferencias en las magnitudes de sismos por escala de medición", #Titulo
        las=2 #Orientación
        )
```

Respuestas práctica 2 b)



Se puede observar que se observan bstante similares los boxplots, pero algo que destaca es que hay escalas en las que hay numerosos outliers, por ejemplo, las escalas MWW, MWC, MWB, etcétera.

Complemento adicional

Filtros múltiples

¿Qué pasa si quisieramos extraer estadísticas para cada una de las 9 escalas? Sería un enorme trabajo filtrar 9 veces. Existe una solución mucho más rápida, con la libería `dplyr`.

Por ejemplo, el siguiente código calcula la media de la magnitud del sismo para cada una de las 9 escalas:

```
library(dplyr) #Carga la librería dplyr (debe estar instalada)

sismos %>% #Utilizamos los datos
  group_by(‘Magnitude Type‘) %>% #Los agrupamos por escala
  summarise(Media=mean(Magnitude)) #Calculamos por grupo la media de la magnitud
# A tibble: 9 x 2
  ‘Magnitude Type‘ Media
  <chr>            <dbl>
1 MB                5.70
2 MD                6.3
3 ML                5.75
4 MS                5.80
5 MW                6.24
6 MWB               5.84
7 MWC               5.86
8 MWR               5.63
9 MWW               6.01
```

Múltiples estadísticas

Y podemos extraer más que solo la media, podemos extraer mediana, mínimo, cantidad de observaciones, etcétera:

```
sismos %>% #Utilizamos los datos
  group_by(`Magnitude Type`) %>% #Los agrupamos por escala
  summarise(Minimo=min(Magnitude), #Calculamos por grupo el minimo de la magnitud
            Mediana=median(Magnitude), #Calculamos por grupo la mediana de la magnitud
            Media=mean(Magnitude), #Calculamos por grupo la media de la magnitud
            Maximo=max(Magnitude), #Calculamos por grupo el maximo de la magnitud
            n=n() #Calculamos por grupo la cantidad de observaciones (en este caso, sismos)
            )
# A tibble: 9 x 6
  `Magnitude Type` Minimo Mediana Media Maximo     n
  <chr>             <dbl>   <dbl> <dbl>  <dbl> <int>
1 MB                  5.5    5.6   5.70    6.7   605
2 MD                  5.8    6.3   6.3     6.8     2
3 ML                  5.5    5.65  5.75    6.2    14
4 MS                  5.5    5.6   5.80    6.7    30
5 MW                  5.5    5.92  6.24    9.1    28
6 MWB                 5.5    5.7   5.84    7.8  1978
7 MWC                 5.5    5.7   5.86    8.3  4523
8 MWR                 5.5    5.5   5.63    6.2    26
9 MWW                 5.5    5.9   6.01    9.1  1983
```

Esto es super útil y rápido!

Explicación