SOC4001 Procesamiento avanzado de bases de datos en R

Tarea 1, respuestas

Ponderación: 12% de la nota final del curso Entrega: Desde el momento de entrega, los estudiantes tiene 1 exacta semana de plazo para completar esta tarea. Formato: Desarrollar esta tarea en un RScript, agregando comentarios cuando sea necesario.

1) Instalar y cargar el paquete (desde el Script) `CarData`.

```
install.packages("carData", repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/6z/_w4wbvf95_bcpp9w2g5nmjr40000gn/T//RtmpfXihlg/downloaded_packages
```

```
library("carData")
```

2) Usa la documentación del paquete `CarData` para identificar los datos correspondientes a "Duncan's Occupational Prestige Data"

3) Carga los datos y crea un objeto que los contenga. Llama tal objeto "datos_duncan".

```
data("Duncan")
datos_duncan <- Duncan
rm(Duncan) # remueve "flotante"
```

4) Muestra las primeras y las últimas 6 observaciones de la base de datos en la consola.

```
head(datos_duncan)
```

```
##            type income education prestige
## accountant prof     62        86       82
## pilot      prof     72        76       83
## architect  prof     75        92       90
## author     prof     55        90       76
## chemist    prof     64        86       90
## minister   prof     21        84       87
```

```
tail(datos_duncan)
```

```
##            type income education prestige
## cook         bc     14        22       16
## soda.clerk   bc     12        30        6
## watchman     bc     17        25       11
## janitor      bc      7        20        8
## policeman    bc     34        47       41
## waiter       bc      8        32       10
```

5) Crea una base de datos que contenga sólo las variables `type` and `prestige` de "datos_duncan". Llama tal objeto "subdatos_duncan". Muestra las dimensiones de la nueva bases de datos.

```
subdatos_duncan <- datos_duncan[,c("type","education","prestige")]
dim(subdatos_duncan)
```

```
## [1] 45  3
```

6) Presenta un resumen estadístico (summary) de las variables en la base de datos.

```
summary(subdatos_duncan)
```

```
##    type      education         prestige
##  bc  :21   Min.   :  7.00   Min.   : 3.00
##  prof:18   1st Qu.: 26.00   1st Qu.:16.00
##  wc  : 6   Median : 45.00   Median :41.00
##            Mean   : 52.56   Mean   :47.69
##            3rd Qu.: 84.00   3rd Qu.:81.00
##            Max.   :100.00   Max.   :97.00
```

7) Chequea la presencia de valores perdidos en la variable "education". Luego, calcula la media de la variable "education" y almacénala en un objeto llamado "education_promedio".

```
is.na(subdatos_duncan$education)
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
education_promedio <- mean(subdatos_duncan$education)
```

8) Crea una nueva variable llamada "educacion_agg". Asigna valor 1 a "educacion_agg" para aquellas observaciones en las cuales la variable "education" toma un valor por sobre la media. Asigna valor 0 a "educacion_agg" para aquellas observaciones en las cuales la variable "education" toma un valor igual o menor a la media.

```
subdatos_duncan$educacion_agg <- NA
subdatos_duncan$educacion_agg[subdatos_duncan$education > education_promedio] <- 1
subdatos_duncan$educacion_agg[subdatos_duncan$education <= education_promedio] <- 0
```

9) Usa un loop para calcular la media y la desviación estándar de la variable "prestige" para las observaciones en cada uno de los niveles de la variable "educacion_agg". No olvides usar el comando `print()` para mostrar los cálculos ejecutados dentro del loop.

```
for (i in 0:1) {
  print(mean(subdatos_duncan$prestige[subdatos_duncan$educacion_agg==i]))
  print(sd(subdatos_duncan$prestige[subdatos_duncan$educacion_agg==i]))
}
```

```
## [1] 26.12
## [1] 20.25323
## [1] 74.65
## [1] 20.09785
```

10) Crea un scatterplot de las variables “prestige” and “education”. Dale un nombre informativo a cada eje.

```
plot(subdatos_duncan$prestige,subdatos_duncan$education, xlab="Prestige", ylab="Education")
```