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")
```

```
## Error in install.packages : Updating loaded packages
```

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

2) Usa la documentación del paquete `CarData` para identificar los datos correspondientes a "Self-Reports of Height and height"

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

```
data("Davis")
datos_davis <- Davis
rm(Davis) # remueve "flotante"
```

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

```
head(datos_davis)
```

```
##   sex weight height repwt repht
## 1   M     77    182    77   180
## 2   F     58    161    51   159
## 3   F     53    161    54   158
## 4   M     68    177    70   175
## 5   F     59    157    59   155
## 6   M     76    170    76   165
```

```
tail(datos_davis)
```

```
##     sex weight height repwt repht
## 195   F     62    164    61   161
## 196   M     74    175    71   175
## 197   M     83    180    80   180
## 198   M     81    175    NA    NA
## 199   M     90    181    91   178
## 200   M     79    177    81   178
```

5) Crea una base de datos que contenga sólo las variables `sex`, `height` y `repht` de "datos_davis". Llama tal objeto "subdatos_davis". Muestra las dimensiones de la nueva bases de datos.

```
subdatos_davis <- datos_davis[,c("sex","height","repht")]
dim(subdatos_davis)
```

```
## [1] 200   3
```

6) Presenta un resumen estadístico (summary) de las variables en "subdatos_davis".

```
summary(subdatos_davis)
```

```
##  sex         height          repht
##  F:112   Min.   : 57.0   Min.   :148.0
##  M: 88   1st Qu.:164.0   1st Qu.:160.5
##          Median :169.5   Median :168.0
##          Mean   :170.0   Mean   :168.5
##          3rd Qu.:177.2   3rd Qu.:175.0
##          Max.   :197.0   Max.   :200.0
##                          NA's   :17
```

7) Crea una variable llamada "diff" que mida la diferencia entre el peso real y el peso reportado por los individuos y añadela a "subdatos_davis".

```
subdatos_davis$diff <- subdatos_davis$height -  subdatos_davis$repht
```

8) Chequea la presencia de valores perdidos en la variable "diff". Luego crea una nueva base de datos que contenga sólo las observaciones con datos completos en todas las variables en "subdatos_davis". Llama este objeto "subdatos_davis_full" y presenta un resumen estadístico (summary) de las variables en "subdatos_davis_full".

```
is.na(subdatos_davis$diff)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [22] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [43] FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE
##  [64] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [106] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [127] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## [148] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
## [169] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE
## [190] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
```

```
subdatos_davis_full <- subdatos_davis[complete.cases(subdatos_davis),]
summary(subdatos_davis_full)
```

```
##  sex         height        repht            diff
##  F:101   Min.   : 57   Min.   :148.0   Min.   :-106.000
##  M: 82   1st Qu.:164   1st Qu.:160.5   1st Qu.:   1.000
```

```
##          Median :169   Median :168.0   Median :   2.000
##          Mean   :170   Mean   :168.5   Mean   :   1.481
##          3rd Qu.:178   3rd Qu.:175.0   3rd Qu.:   3.000
##          Max.   :197   Max.   :200.0   Max.   :  10.000
```

8) Crea una nueva variable llamada "sex_num". Asigna valor 1 a "sex_num" para aquellas observaciones en las cuales la variable "sex" toma valor "F" (mujer). Asigna valor 0 a "sex_num" para aquellas observaciones en las cuales la variable "sex" toma un valor "M" (hombre).

```
subdatos_davis_full$sex_num[subdatos_davis_full$sex == "F"] <- 1
subdatos_davis_full$sex_num[subdatos_davis_full$sex == "M"] <- 0
```

9) Usa un loop para calcular la media de la variable "diff" para las observaciones en cada uno de los niveles de la variable "sex" (es decir, para hombres y mujeres). No olvides usar el comando `print()` para mostrar los cálculos ejecutados dentro del loop.

```
for (i in c("F","M")) {
  print(i)
  print(mean(subdatos_davis_full$diff[subdatos_davis_full$sex==i]))
}
```

```
## [1] "F"
## [1] 1.257426
## [1] "M"
## [1] 1.756098
```

10) Explica MUY brevemente el resultado obtenidos en la pregunta anterior.

En promedio, tanto mujeres como hombres miden más de lo que reportan, especialmente los hombres.