SOC4001 Procesamiento avanzado de bases de datos en R

Tarea 2, respuestas

Ponderación: 12% de la nota final del curso

Formato: Desarrollar esta tarea en un RScript, agregando comentarios cuando sea necesario.

1) Carga la base de datos "Chile" del paquete `carData` y crea un objeto que los contenga los datos. Llama tal objeto "datos_chile". Carga la librería `tidyverse` y ejecuta la siguientes operaciones usando las herramientas contenidas de `tidyverse`:

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data("Chile")
datos_chile <- Chile
rm(Chile) # remueve "flotante"

datos_chile %>% glimpse()
```

```
## Rows: 2,700
## Columns: 8
## $ region     <fct> N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N,~
## $ population <int> 175000, 175000, 175000, 175000, 175000, 175000, 175000, 175~
## $ sex        <fct> M, M, F, F, F, F, M, F, F, M, M, M, F, F, M, M, F, M, M, F,~
## $ age        <int> 65, 29, 38, 49, 23, 28, 26, 24, 41, 41, 64, 19, 27, 46, 36,~
## $ education  <fct> P, PS, P, P, S, P, PS, S, P, P, P, S, PS, S, PS, S, PS, S, ~
## $ income     <int> 35000, 7500, 15000, 35000, 35000, 7500, 35000, 15000, 15000~
## $ statusquo  <dbl> 1.00820, -1.29617, 1.23072, -1.03163, -1.10496, -1.04685, -~
## $ vote       <fct> Y, N, Y, N, N, N, N, N, U, N, Y, U, Y, Y, NA, A, N, U, Y, U~
```

2) Añade a "datos_chile" un variable llamada "year" con valor 1988 en todas las filas

```
datos_chile <- datos_chile %>% mutate(year = 1988)
```

3) Calcula el año de nacimiento de cada individuo. Añade a "datos_chile" un variable llamada "birthyear" que contenga esta información

```
datos_chile <- datos_chile %>% mutate(birthyear = year - age)
```

4) Usando la función `if_else()` añade a "datos_chile" un variable llamada "vote_no" que tome valor 1 si la persona declara que votará por el No y valor 0 en cualquier otra caso.

```
datos_chile <-  datos_chile %>% mutate(vote_no = if_else(vote=="N",1,0))
```

5) Usando la función `case_when()` añade a "datos_chile" un variable llamada "cohort73" que tome valor 1 si la persona tenía 18 año o más el año del golpe de estado (1973) y valor 0 si tenía menos de 18. Trata las observaciones que no cumplan ninguna de estas condiciones como valores perdidos.

```
datos_chile <-  datos_chile %>% mutate(cohort73 = case_when(birthyear <= (1973 - 18) ~ 1,
                            birthyear > (1973 - 18) ~ 0)
                          )
```

6) Usando la función `group_by()` añade a "datos_chile" un variable llamada "no_by_groups" que contenga el promedio de la variable "vote_no" por región, nivel educacional y cohorte (cohort73).

```
datos_chile <-  datos_chile %>% group_by(region,education,cohort73) %>%
                mutate(no_by_groups = mean(vote_no, na.rm = T))
```

```
datos_chile %>% select(no_by_groups) %>% glimpse()
```

```
## Adding missing grouping variables: `region`, `education`, `cohort73`
```

```
## Rows: 2,700
## Columns: 4
## Groups: region, education, cohort73 [35]
## $ region       <fct> N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, N, ~
## $ education    <fct> P, PS, P, P, S, P, PS, S, P, P, P, S, PS, S, PS, S, PS, S~
## $ cohort73     <dbl> 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, ~
## $ no_by_groups <dbl> 0.2020202, 0.5312500, 0.2020202, 0.2020202, 0.3939394, 0.~
```