OPSO79-1-UCSH2021

Revisión de prueba 1

01/10/2021

Revisión prueba 1.

Apartado práctico.

Módulo II

Pregunta 2.1

Si deseo exportar y compartir un microdato, ¿que tipo de archivo resultaría más pertinente en R?, ¿.rmd, .html, .rds, .r o .Rhistory? Solo escriba el tipo de archivo.

`.rds`

Pregunta 2.2

Intente explicar en palabras simples, para alguien con escaso manejo de R, en qué consiste un paquete y cuál sería la diferencia con una data frame (basta con un párrafo)

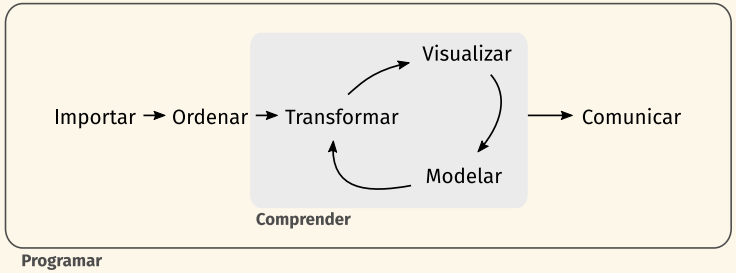
Los paquetes en R son colecciones de funciones y conjunto de datos desarrollados por la comunidad. Estos incrementan la potencialidad de R mejorando las funcionalidades base en R, o añadiendo nuevas. Los paquetes, en tanto pueden contener conjuntos de datos, pueden almacenar objetos del tipo *data frame*.

Los data frames son estructuras de datos de dos dimensiones (rectangulares) que pueden contener datos de diferentes tipos (carácter, numérico, enteros, etc.).

Módulo II

Pregunta 2.3

En cuál etapa del esquema de la ciencia de datos de Hadley Wickham se ubicarían los paquetes `haven` y `dplyr`. Indique brevemente las razones (no más de dos líneas para cada paquete)



haven para la importación de datos

dplyr para la transformación de datos. Requiere que ya existan datos cargados y que estos se encuentren en formato rectangular y ordenado (formato `tidy`)

Módulo II

Pregunta 2.4

¿Por qué motivos un usuario de R optaría por trabajar con `R Project` y no simplemente abriendo R y luego un script en el cuál codificar? (no más de un párrafo)

R trabaja con **directorios de trabajo**. Aquí es donde R busca los archivos que le pedimos que lea y donde colocará todos los archivos que le pidamos que guarde. Con los RProject podemos definir cuál será el directorio de trabajo que utilizaremos. Esto asegura un fácil acceso a los archivos (data, imágenes) que son utilizados en un proyecto.

Además, permite asegurar la **reproducibilidad** del código. Esto es, que mi yo del futuro o cualquier otra personas pueda ejecutar el código que escribí desde otra computadora y llegar al mismo resultado. Esto se logra evitando escribir rutas absolutas cuando importamos información a R.

Módulo III

La Encuesta Suplementaria de Ingresos (ESI), es un módulo complementario que se aplica dentro de la Encuesta Nacional de Empleo (ENE).

La página web de la encuesta es la siguiente: https://www.ine.cl/estadisticas/sociales/ingresos-y-gastos/encuesta-suplementaria-de-ingresos

Descargue el microdato del año 2020 (ESI 2020 - Personas), cargue la data en R y responda.

Cargar paquetes

```
library(haven)
library(dplyr)
```

Cargar data

```
esi2020 <- read_sav("data/esi-2020---personas_s.sav")
```

Módulo III

Pregunta 3.1

¿Cuántas personas hay en la región de Aysén?

Usar `table(esi2020$region)` era la opción más simple, pero había que buscar número asociado a región en libro de variables de la ESI.

```
table(esi2020$region)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12
## 2139 2657 2513 3774 9333 4720 5388 8474 3869 5348 1429 1548 123
##   14   15   16
## 3062 2134 3162
```

La opción aprovechando las etiquetas:

```
sjmisc::frq(esi2020$region)
```

```
##
## Región (x) <numeric>
## # total N=71935  valid N=71935  mean=8.61  sd=4.07
##
## Value |                  Label |     N | Raw % | Valid % | Cum. %
## --------------------------------------------------------------------
##     1 |               Tarapacá |  2139 |  2.97 |    2.97 |   2.97
##     2 |             Antofagasta |  2657 |  3.69 |    3.69 |   6.67
##     3 |                Atacama |  2513 |  3.49 |    3.49 |  10.16
##     4 |               Coquimbo |  3774 |  5.25 |    5.25 |  15.41
##     5 |             Valparaíso |  9333 | 12.97 |   12.97 |  28.38
##     6 |              O'Higgins |  4720 |  6.56 |    6.56 |  34.94
##     7 |                 Maule |  5388 |  7.49 |    7.49 |  42.43
##     8 |                Biobío |  8474 | 11.78 |   11.78 |  54.21
##     9 |          La Araucanía |  3869 |  5.38 |    5.38 |  59.59
##    10 |             Los Lagos |  5348 |  7.43 |    7.43 |  67.03
##    11 |                 Aysén |  1429 |  1.99 |    1.99 |  69.01
##    12 |            Magallanes |  1548 |  2.15 |    2.15 |  71.16
##    13 |          Metropolitana | 12385 | 17.22 |   17.22 |  88.38
##    14 |              Los Ríos |  3062 |  4.26 |    4.26 |  92.64
##    15 |     Arica y Parinacota |  2134 |  2.97 |    2.97 |  95.60
##    16 |                 Ñuble |  3162 |  4.40 |    4.40 | 100.00
##    99 | Región no identificada |     0 |  0.00 |    0.00 | 100.00
##  <NA> |                  <NA> |     0 |  0.00 |    <NA> |   <NA>
```

En la región de Aysén hay 1429 personas en la muestra

Módulo III

Pregunta 3.2

¿Cuántas personas en Chile tienen ingresos de más de 1 millón de pesos por concepto de sueldos y salarios? (ocupar variable `ing_t_d`)

Lo que no había que hacer (tabular y luego sumar)

```
table(esi2020)
```

La idea es crear una nueva variable con condición, luego filtrar y sumar a las personas

```
esi2020<- mutate(esi2020,filtro_mas_de_1M=if_else(ing_t_d>1000000,1,0))
table(esi2020$filtro_mas_de_1M)
```

```
##
##     0     1
## 69136  2799
```

2799 personas de la muestra ganan más de 1 millon de pesos

Módulo III

Pregunta 3.3

¿Cuánto es lo máximo, mínimo y el promedio de ingresos que gana por concepto de sueldos y salarios una persona en Chile?, ¿Le parecen razonables los resultados?

```
summary(esi2020$ing_t_d)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0  162616   80105 9312239
```

La mitad o más de la población gana $0 y en promedio las personas ganan $162.616.

No hace sentido.

Módulo III

Pregunta 3.4

Repita el ejercicio de la pregunta 3.3 pero sin considerar a las personas que registran cero pesos por sueldos y salarios.

Se quitan aquellos casos que ganan menos de 0 pesos.

```
esi2020_2 <- select(filter(esi2020,ing_t_d>0),ing_t_d)
```

Se vuelve a hacer un summary

```
summary(esi2020_2$ing_t_d)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5007  320421  450000  639851  750987 9312239
```

Ahora la mediana es $450.000 y la media $639.851. Valores similares a los reportados en resultados oficiales.

De todas formas los valores mínimos y máximos son extraños.

Módulo III

Pregunta 3.5

¿Cuál es la brecha o diferencia de ingresos totales entre hombres y mujeres?

Para hacer la comparación quite los valores cero de la variable de ingresos.

De esta forma vemos las etiquetas de la variable sexo:

```
sjmisc::frq(esi2020$sexo)
```

```
##
## Sexo (x) <numeric>
## # total N=71935  valid N=71935  mean=1.53  sd=0.50
##
## Value |  Label |     N | Raw % | Valid % | Cum. %
## ---------------------------------------------------
##     1 | Hombre | 33830 | 47.03 |   47.03 |  47.03
##     2 |  Mujer | 38105 | 52.97 |   52.97 | 100.00
##  <NA> |   <NA> |     0 |  0.00 |    <NA> |   <NA>
```

Módulo 3

Creamos 2 sub conjuntos de datos. Uno solo con mujeres, otro solo con hombres.

```
mujeres <- select(filter(esi2020,ing_mon_sb>0 & sexo==2),ing_mon_sb)
hombres <- select(filter(esi2020,ing_mon_sb>0 & sexo==1),ing_mon_sb)
```

Con media:

```
mean(mujeres$ing_mon_sb)-mean(hombres$ing_mon_sb)
```

```
## [1] -179288.7
```

Con mediana:

```
median(mujeres$ing_mon_sb)-median(hombres$ing_mon_sb)
```

```
## [1] -125850.7
```

Conclusión general: las mujeres tienen reciben menos ingresos que los hombres.

Recursos web utilizados

Xaringan: Presentation Ninja, de Yihui Xie. Para generar esta presentación.

Bibliografía utilizada

Wickham, H. (2021). *R Para Ciencia de Datos*.