OPSO79-1-UCSH2021

Metodologías de investigación social y desafíos de la sociología. Bloque práctico.

20-08-2021

Repaso lenguaje R y comentarios sobre texto ciencia de datos

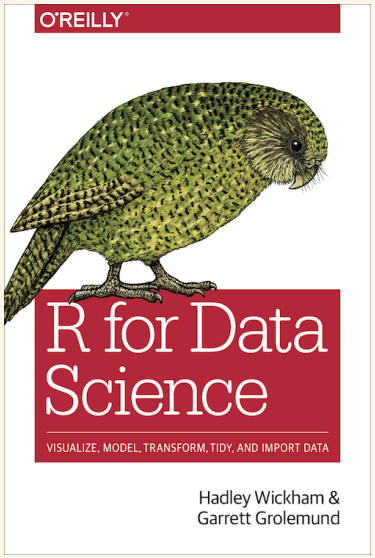
R repaso

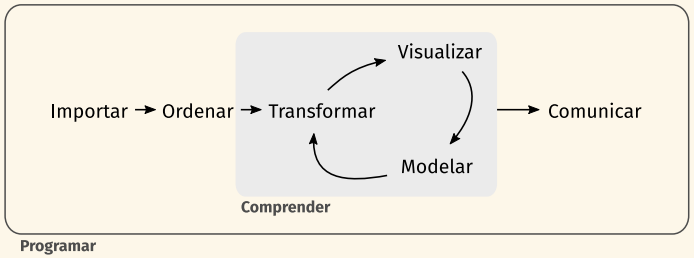
- R lenguaje de programación para el análisis estadístico (archivos `.R`)
- RStudio interfaz "amigable" (IDE)
- Markdown lenguaje de texto plano y de marcado
- RMarkdown integración de los dos lenguajes (archivos `.rmd`)
- RMarkdown permite integrar en un mismo documento el análisis de los datos y su interpretación.
- Documentos en RMarkdown son reproducibles
- Para agregar imagen: `![](rutadelaimagen.png)`


Rmarkdown
TEXT. CODE. OUTPUT.
(GET IT TOGETHER, PEOPLE.)
#HEADERS
CODE
TEXT
PANDOC
KNITR
FAIRY DUST
RECIPE
1. ADD TEXT
2. ADD CODE
3. KNIT
4. (magic)
5. CELEBRATE PERCEIVED WIZARDRY
rmarkdown
knitr
pandoc
.Rmd
.md

Ciencia de datos

El modelo de la ciencia de datos según Wickham (2021):





Ciencia de datos

La pregunta sobre cuáles datos importar y con qué objetivo debe ser central para la sociología.

La comunicación no se limita a presentar el dato. Hay que interpretarlo, ponerlo en discusión teórica e intentar dar respuesta a pregunta inicial.

Debemos ser consciente de cómo los datos son producidos y distribuidos (o no distribuidos). Proceso de producción y análisis estadístico por instituciones.

R

Importar, ordenar y transformar los datos no es sencillo.

Visualizar los datos suele ser la parte más sencilla y cercana.

"Una buena visualización te mostrará cosas que no esperabas o hará surgir nuevas preguntas acerca de los datos." Wickham (2021).

Trabajaremos con una data frame que se carga fácilmente, está ordenada y transformada según nuestro interés inicial.

Antes, revisemos cosas básicas de R y algunos conceptos.

Lenguaje R

Vectores

Es el objeto más básico en R.

Un vector es una forma de almacenar datos que permite contener una serie de valores del mismo tipo.

Simples de solo 1 elemento:

```
a<-7
a
```

```
## [1] 7
```

Cosas más complejas con la función `c()` para concatenar elementos:

```
a<-c(7,8,9,5,7,9)
a
```

```
## [1] 7 8 9 5 7 9
```

Vectores

Podemos crear un vector con los nombres del curso:

```
nombres<-c("ISIDORA", "ALEJANDRA", "NOEMÍ", "LESLIE","MARCO", "JAVIER",
           "DANTE","VALENTINA", "DIANA" ,  "ANDRÉS",   "BRUNO", "JAVIERA
           "JAVIERA" , "PAOLA", "ALONSO" ,  "THABATA", "JENNIFER", "FRAN
           "MATÍAS",  "VALENTINA", "GERALDINE","ALEJANDRA" , "VALENTINA"
           "FRANCISCO", "DIEGO", "MATIAS", "CAMILA" , "CARLA",
           "JAVIER")
```

Ver el vector:

```
nombres
```

```
##  [1] "ISIDORA"   "ALEJANDRA" "NOEMÍ"     "LESLIE"    "MARCO"     "JAVIER"
##  [7] "DANTE"     "VALENTINA" "DIANA"     "ANDRÉS"    "BRUNO"     "JAVIERA"
## [13] "JAVIERA"   "PAOLA"     "ALONSO"    "THABATA"   "JENNIFER"  "FRANCISCO
## [19] "MATÍAS"    "VALENTINA" "GERALDINE" "ALEJANDRA" "VALENTINA" "FRANCISCO
## [25] "DIEGO"     "MATIAS"    "CAMILA"    "CARLA"     "JAVIER"
```

Vectores e indexación

Y luego seleccionar la posición 8.

```
nombres[8]
```

```
## [1] "VALENTINA"
```

O también las posiciones 8, 12 y 15

```
nombres[c(8,12,13)]   ## necesitamos c()
```

```
## [1] "VALENTINA" "JAVIERA"   "JAVIERA"
```

O de la 15 a la 17

```
nombres[c(15:17)]
```

```
## [1] "ALONSO"   "THABATA"  "JENNIFER"
```

Vectores

Creamos dos vectores de distintos tipo.

Existen 5 tipos de vectores en R:

```
character <- c("gato", "perro")
numeric <- c(8, 15.9) # reales o decimales
integer <-  c(2L, 4L) # L indica que son enteros
logical <- c(TRUE, FALSE, TRUE)
complex <- 3 + 4i # complejos
```

Podemos saber su tipo, preguntándole a R:

```
class(nombres)
```

```
## [1] "character"
```

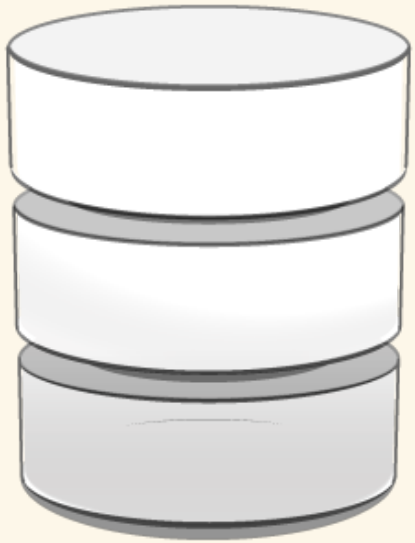
```
class(a)
```

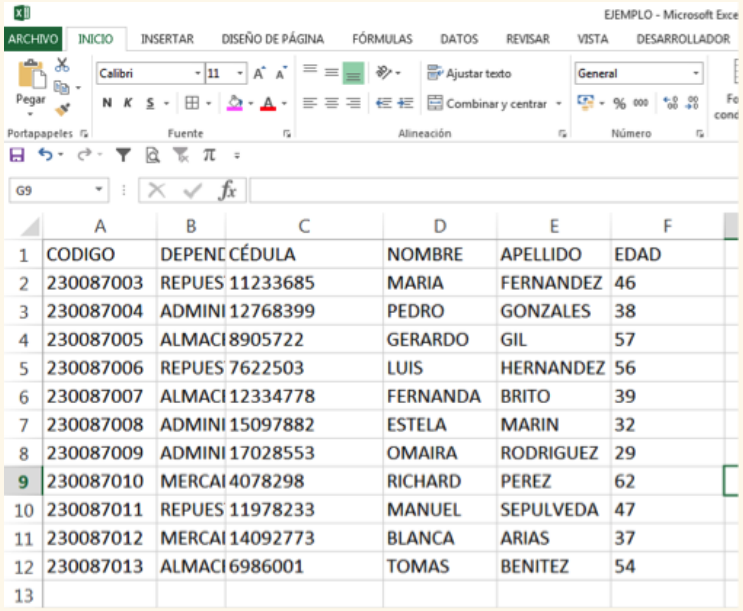
```
## [1] "numeric"
```

Data frames

Las `data frames` son lo que en SPSS llamamos "bases de datos".

Técnicamente son tablas, estructuras rectangulares de observaciones y variables.





	A	B	C	D	E	F
1	CODIGO	DEPENI	CÉDULA	NOMBRE	APELLIDO	EDAD
2	230087003	REPUES	11233685	MARIA	FERNANDEZ	46
3	230087004	ADMINI	12768399	PEDRO	GONZALES	38
4	230087005	ALMACI	8905722	GERARDO	GIL	57
5	230087006	REPUES	7622503	LUIS	HERNANDEZ	56
6	230087007	ALMACI	12334778	FERNANDA	BRITO	39
7	230087008	ADMINI	15097882	ESTELA	MARIN	32
8	230087009	ADMINI	17028553	OMAIRA	RODRIGUEZ	29
9	230087010	MERCAI	4078298	RICHARD	PEREZ	62
10	230087011	REPUES	11978233	MANUEL	SEPULVEDA	47
11	230087012	MERCAI	14092773	BLANCA	ARIAS	37
12	230087013	ALMACI	6986001	TOMAS	BENITEZ	54
13						

Paquetes

"Incluyen funciones reutilizables, la documentación que describe cómo usarlas y datos de muestra" Wickham (2021).

Estas funciones, datos y documentación permiten extender las capacidades de R base.

Cualquier persona puede hacer un paquete, es código compartido para que terceros puedan usarlo.

Los servidores de CRAN almacenan los paquetes que ya han sido probados y autorizados por la comunidad R.

Para usar paquetes debemos instalarlos con `install.packages()` y cargarlos con `library()`

Paquetes


Shiny
RStudio
ggplot2
dplyr
rmarkdown
knitr
lubridate
stringr
devtools
tidyr
%>%
Ceci n'est pas un pipe.
packrat
ggvis
readr

Paquetes guaguas

Datos sobre nombres de guaguas (bebés) registrados en Chile entre 1920 y 2020, según el Servicio de Registro Civil e Identificación.

README.md

guaguas

CRAN 0.1.0 build passing

Datos de nombres de guaguas (bebés) registrados en Chile entre 1920 y 2019, según el Servicio de Registro Civil e Identificación. Incluye todos los nombres con al menos 15 ocurrencias. Este *dataset* permite explorar tendencias en los nombres registrados durante el último siglo y puede utilizarse como fuente de datos de práctica para enseñar/aprender R.

Es un paquete con datos (`data frame`), no con funciones.

```
#install.packages("guaguas")  ## Instalar paquete (solo una vez)
library(guaguas)              ## Cargamos el paquete (cada vez que usem
```

Paquetes guaguas

```
guaguas <- guaguas   ## cargamos en nuestro ambiente guaguas como "guagu
```

¿Cuántas columnas y cuantas filas tiene la base de datos?

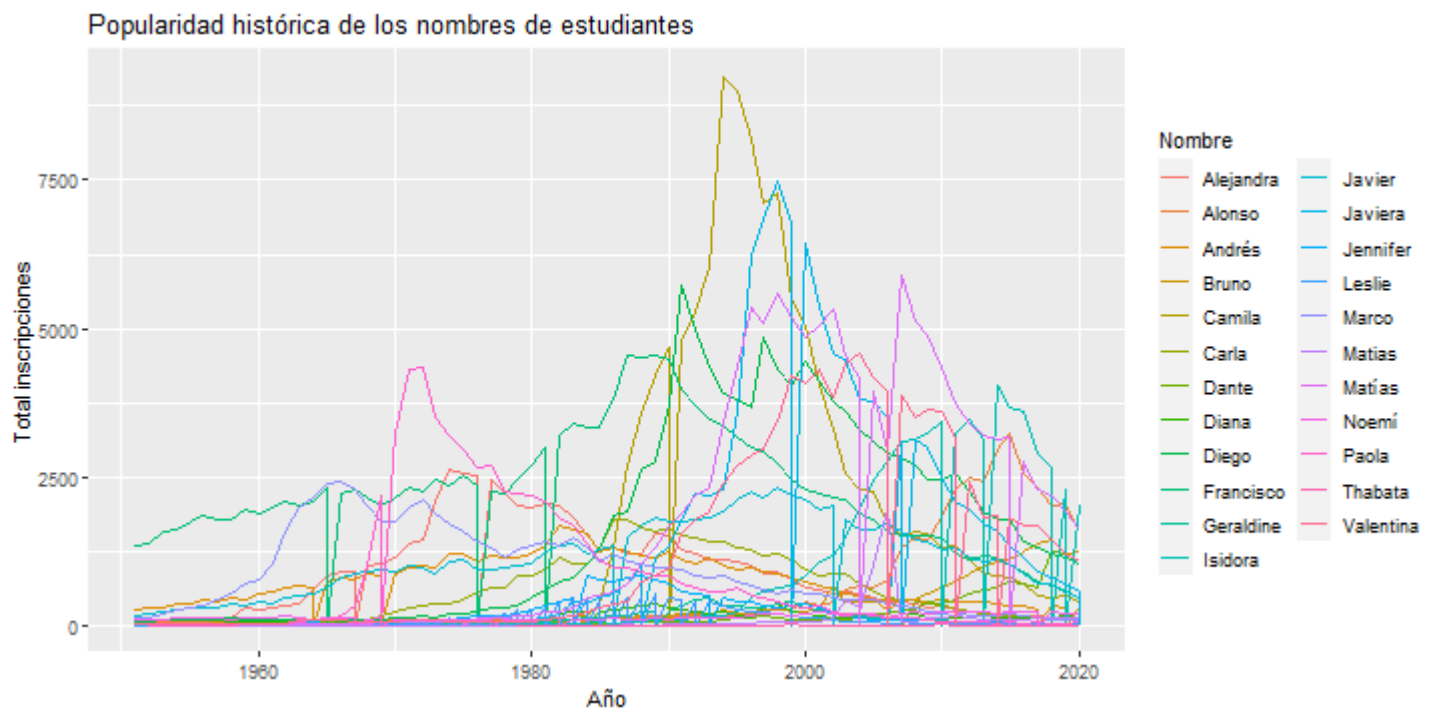
```
dim(guaguas)
```

```
## [1] 845400      5
```

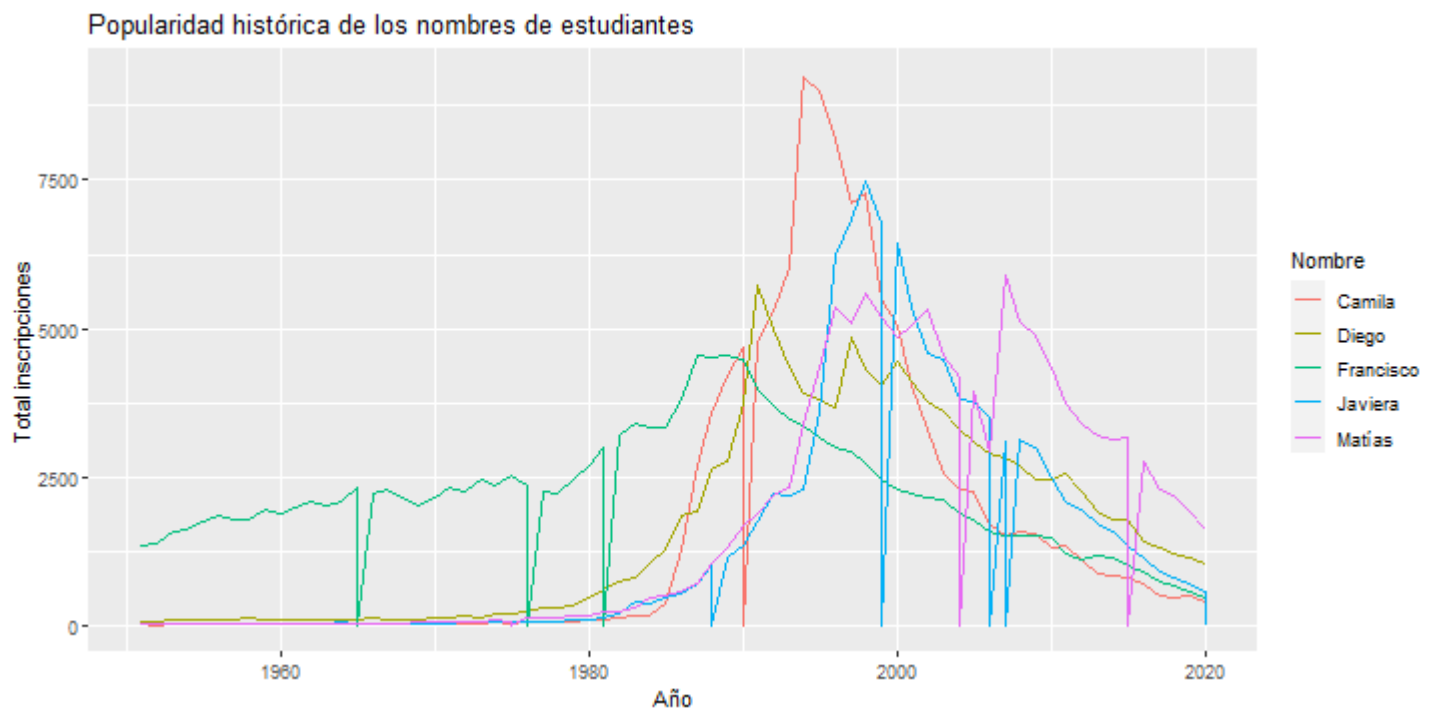
```
head(guaguas)
```

```
## # A tibble: 6 x 5
##    anio nombre sexo      n proporcion
##   <dbl> <chr>  <chr> <dbl>      <dbl>
## 1  1920 María  F      2130     0.104
## 2  1920 José   M       984     0.0483
## 3  1920 Juan   M       636     0.0312
## 4  1920 Luis   M       631     0.0310
## 5  1920 Rosa   F       426     0.0209
## 6  1920 Ana    F       340     0.0167
```

Gráfico con guaguas



Solo los más populares



Funciones elementales

Contar el número de observaciones

```
nrow(guaguas)
```

```
## [1] 845400
```

Conocer el nombre de las variables

```
names(guaguas)
```

```
## [1] "anio"       "nombre"     "sexo"       "n"          "proporcion"
```

Seleccionar primera fila

```
guaguas[1,]
```

```
## # A tibble: 1 x 5
##    anio nombre sexo      n proporcion
##   <dbl> <chr>  <chr> <dbl>      <dbl>
## 1  1920 María  F      2130      0.104
```

Funciones elementales

Seleccionar una variable

```
guaguas[,2]
```

```
## # A tibble: 845,400 x 1
##    nombre
##    <chr>
##  1 María
##  2 José
##  3 Juan
##  4 Luis
##  5 Rosa
##  6 Ana
##  7 Manuel
##  8 Olga
##  9 Carlos
## 10 Pedro
## # ... with 845,390 more rows
```

Funciones elementales

Tabular una variable

```
table(guaguas[,3])
```

```
##
##      F      M
## 523623 321777
```

```
prop.table(table(guaguas[,3]))
```

```
##
##        F        M
## 0.619379 0.380621
```

Sacar el porcentaje de un tabulado

```
prop.table(table(guaguas[,3]))*100
```

```
##
##       F       M
```

Funciones elementales

Filtrar según valor (por ejemplo, para encontrar un nombre)

La lógica es similar a encontrar una fila

```
guaguas[500,]
```

```
## # A tibble: 1 x 5
##    anio nombre  sexo      n proporcion
##   <dbl> <chr>   <chr> <dbl>      <dbl>
## 1  1920 Bartola F         4   0.000196
```

Pero en vez seleccionar la fila 500, queremos seleccionar según una condición

```
guaguas$nombre=="Nicolás"
```

```
guaguas[guaguas$nombre=="Nicolás",]
```

```
## # A tibble: 6 x 5
##    anio nombre  sexo      n proporcion
##   <dbl> <chr>   <chr> <dbl>      <dbl>
## 1  1920 Nicolás M        17   0.000834
```

Funciones elementales

También se pueden sacar sumas

```
sum(guaguas[,4])
```

```
## [1] 21951078
```

O promedios (otra forma de seleccionar variables)

```
mean(guaguas$n)
```

```
## [1] 25.96532
```

Y mediana:

```
median(guaguas$n)
```

```
## [1] 1
```

Ejercicios para practicar

- ¿Cuantas guaguas se inscribieron en 1920 con tu nombre?
- ¿Cuantas guaguas se inscribieron el año en que naciste con tu nombre?
- Crea una nueva data frame que se llame como tú y que solo contenga inscripciones de tu nombre
- ¿Cuantas observaciones tiene esta nueva data frame?
- ¿Cuántas mujeres y hombres hay en esta nueva data frame?
- Crea una nueva data que solo tenga datos de 2020 (llamala "pandemia")
- ¿Cuánto suma la variable "proporción" de "pandemia"?
- **Desafío**: el nombre más popular en 2020 (mayor N)

Bibliografía utilizada

Wickham, H. (2021). *R Para Ciencia de Datos*. URL: https://es.r4ds.hadley.nz/.

Quiroga, R. (2021). guaguas: Nombres Inscritos en Chile (1920 - 2020). R package version 0.2.0. https://CRAN.R-project.org/package=guaguas

Para seguir aprendiendo

La trágica y heroica biografía de uno de los creadores de Markdown y CC (documental).

El paquete guaguas y los nombres Salvador, Augusto, de Romané y de los Backstreet Boys (Viñeta).