Taller 2: Exploración y validación de datos

José Daniel Conejeros

Magíster(c) Sociología UC - Licenciado en Ciencias Sociales

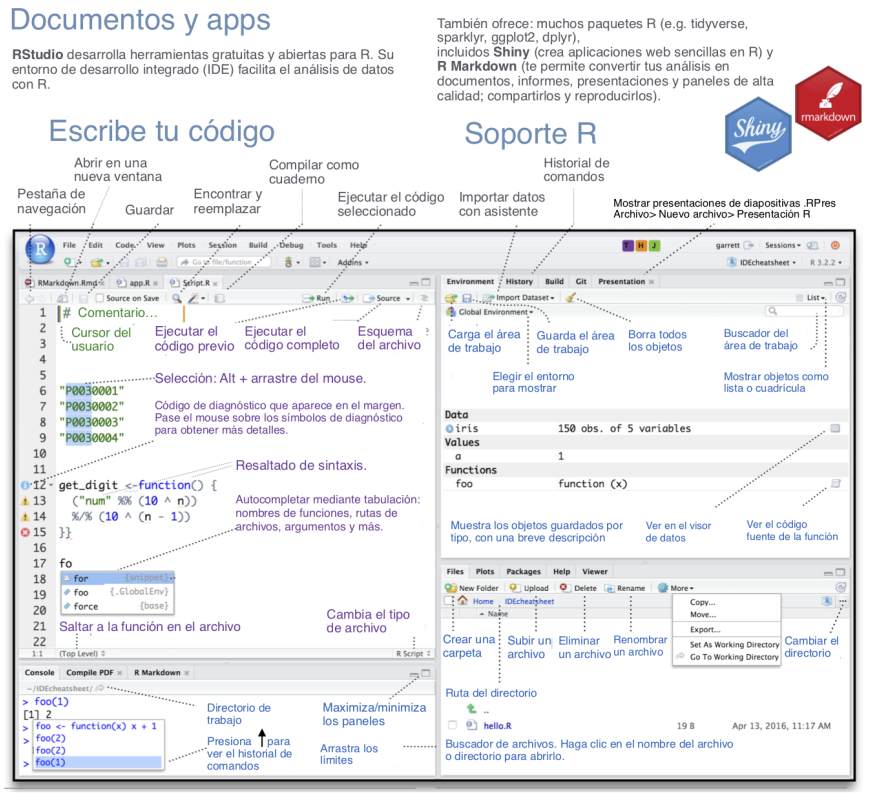
Junio 2020

Código del taller jdconejeros@uc.cl JDConejeros

joseconejerosp Jose_Conejeros

¿Qué hemos visto hasta el momento?

- Instalación de `R` y `RStudio`
- Explorar espacios de trabajo
- Comprender la lógica de objetos
- Explorar vectores, listas y matrices
- Estructura de funciones y argumentos
- Instalar y cargar librerías/paquetes

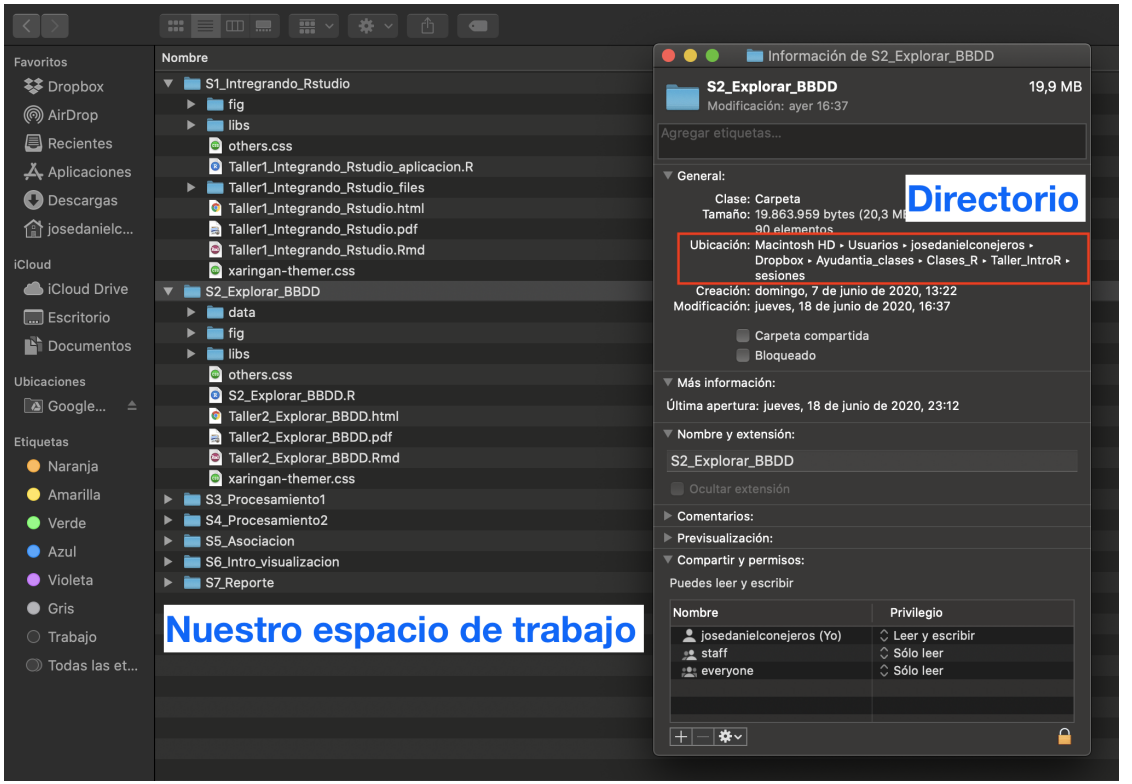


¿Alguna pregunta?

Hoy: Definir nuestro espacio de trabajo

```r
getwd() # Directorio de trabajo actual

setwd("ruta") # Establecer directorio de trabajo
```



También podemos generar proyectos de trabajo

Estructura de una base de datos

Las filas representan las observaciones - Las columnas representan las variables

```
id ← c(1, 2, 2, 4, 4)
edad ← c(23, 45, 67, 201)
sexo ← c(1, 0, 1, 0, NA)
peso ← c(80, 60, 70, 50, 55)
altura ← c(180, 160, 200, 140, 300)

data ← as.data.frame(cbind(id, edad, sexo,
                           peso, altura))
data ← as.data.frame(data)
data
```

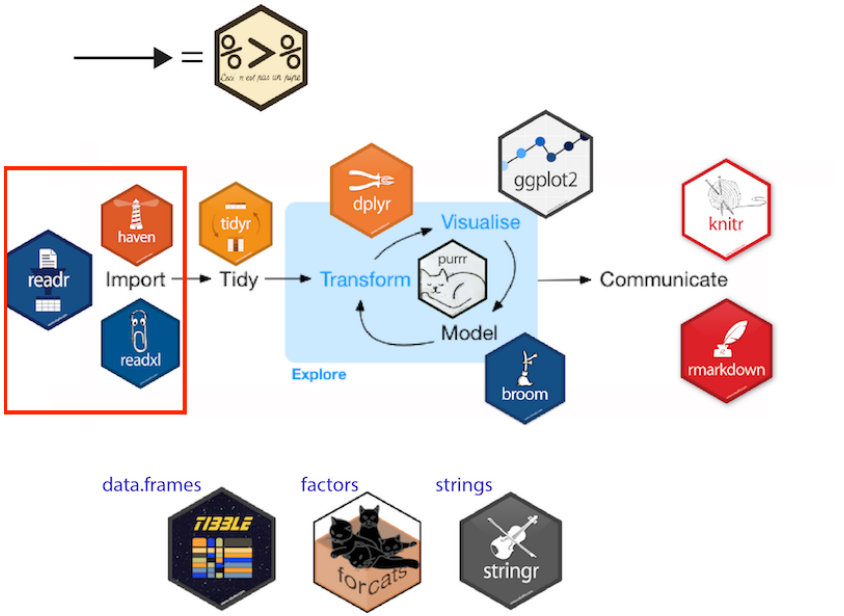
```
##   id edad sexo peso altura
## 1  1   23    1   80    180
## 2  2   45    0   60    160
## 3  2   67    1   70    200
## 4  4  201    0   50    140
## 5  4   23   NA   55    300
```

Dim	Homogeneous	Heterogeneous
1d	Atomic Vector	Matrix
2d	Matrix	Data - Frame
nd	Array	

Mayor detalle pueden revisar el siguiente enlace

Importar una base de datos

```
install.packages(c("haven", "readxl", "readr"))
library(haven)
library(readxl)
library(readr)
```



Mayor detalle del paquete `readr` aquí

Mayor detalle del paquete `haven` aquí

Explorar nuestra base de datos


```
dim(data)   # Observaciones y variables
```

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

```
names(data) # Nombre de nuestras variables
```

```
## [1] "id"     "edad"   "sexo"   "peso"   "altura"
```

```
str(data)   # Visor de nuestras variables
```

```
## 'data.frame':    5 obs. of  5 variables:
##  $ id    : num  1 2 2 4 4
##  $ edad  : num  23 45 67 201 23
##  $ sexo  : num  1 0 1 0 NA
##  $ peso  : num  80 60 70 50 55
##  $ altura: num  180 160 200 140 300
```

Explorar nuestra base de datos


```
head(data)  # Primeras 6 observaciones
```

```
##   id edad sexo peso altura
## 1  1   23    1   80    180
## 2  2   45    0   60    160
## 3  2   67    1   70    200
## 4  4  201    0   50    140
## 5  4   23   NA   55    300
```

```
tail(data)  # Últimas 6 observaciones
```

```
##   id edad sexo peso altura
## 1  1   23    1   80    180
## 2  2   45    0   60    160
## 3  2   67    1   70    200
## 4  4  201    0   50    140
## 5  4   23   NA   55    300
```

Explorar nuestra base de datos


```
summary(data)

library(skimr)
skim(data)
```

skim_type	skim_variable	n_missing	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
numeric	id	0	2.6	1.3416408	1	2	2.0	4	4	▃▇▁▁▇
numeric	edad	0	71.8	74.4929527	23	23	45.0	67	201	▇▂▁▁▁
numeric	sexo	1	0.5	0.5773503	0	0	0.5	1	1	▇▁▁▁▇
numeric	peso	0	63.0	12.0415946	50	55	60.0	70	80	▇▂▂▂▂
numeric	altura	0	196.0	62.2896460	140	160	180.0	200	300	▇▇▁▁▂

¿Qué podemos ver de nuestras variables?

Validación de una base de datos

- **Identificar observaciones duplicadas**

```
# Nuestros casos únicos
unique(data)
```

```
##   id edad sexo peso altura
## 1  1   23    1   80    180
## 2  2   45    0   60    160
## 3  2   67    1   70    200
## 4  4  201    0   50    140
## 5  4   23   NA   55    300
```

```
# Nuestros casos únicos
unique(data$id)
```

```
## [1] 1 2 4
```

```
# Nuestros casos duplicados
duplicated(data$id)
```

```
## [1] FALSE FALSE  TRUE FALSE  TRUE
```

Validación de una base de datos

Identificar observaciones duplicadas

```
# Nuestros casos duplicados
data$caso ← duplicated(data$id)
# Resumen de duplicados
table(data$caso)
```

```
##
## FALSE  TRUE
##     3     2
```

```
# Nuestros casos duplicados
data[duplicated(data$id),]
```

```
##   id edad sexo peso altura caso
## 3  2   67    1   70    200 TRUE
## 5  4   23   NA   55    300 TRUE
```

Validación de una base de datos

Identificar observaciones duplicadas

```
# Quitar los casos duplicados
data2 ← data[unique(data$id),]
data2
```

```
##   id edad sexo peso altura  caso
## 1  1   23    1   80    180 FALSE
## 2  2   45    0   60    160 FALSE
## 4  4  201    0   50    140 FALSE
```

Revisar valores fuera de rango

```
# Mínimo
min(data$edad)
```

```
## [1] 23
```

```
# Máximo
max(data$edad)
```

```
## [1] 201
```

Validación de una base de datos

Revisar valores fuera de rango

```
# Tabla de valores
table(data$edad)
```

```
##
##  23  45  67 201
##   2   1   1   1
```

```
table(data$sexo)
```

```
##
## 0 1
## 2 2
```

```
# Descriptivos
summary(data$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23.0    23.0    45.0    71.8    67.0   201.0
```

Validación de una base de datos

Explorar codificación de los casos pérdidos

```
# Revisar casos perdidos
table(data$sexo, useNA = "ifany")
```

```
##
##    0    1 <NA>
##    2    2    1
```

```
# Mostrar los casos pérdidos
is.na(data)
```

```
##         id  edad  sexo  peso altura  caso
## [1,] FALSE FALSE FALSE FALSE  FALSE FALSE
## [2,] FALSE FALSE FALSE FALSE  FALSE FALSE
## [3,] FALSE FALSE FALSE FALSE  FALSE FALSE
## [4,] FALSE FALSE FALSE FALSE  FALSE FALSE
## [5,] FALSE FALSE  TRUE FALSE  FALSE FALSE
```

Validación de una base de datos

Explorar codificación de los casos pérdidos

```
which(is.na(data$sexo)) # Indica la fila
```

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

```
# Eliminar casos perdidos
data3 ← na.omit(data)
data3
```

```
##   id edad sexo peso altura  caso
## 1  1   23    1   80    180 FALSE
## 2  2   45    0   60    160 FALSE
## 3  2   67    1   70    200  TRUE
## 4  4  201    0   50    140 FALSE
```

Validación de una base de datos

Explorar codificación de los casos pérdidos

```
# Recodificar casos perdidos (imputar)
data4 ← data
data4$sexo[is.na(data$sexo)] ← 1
data
```

```
##   id edad sexo peso altura  caso
## 1  1   23    1   80    180 FALSE
## 2  2   45    0   60    160 FALSE
## 3  2   67    1   70    200  TRUE
## 4  4  201    0   50    140 FALSE
## 5  4   23   NA   55    300  TRUE
```

```
data4
```

```
##   id edad sexo peso altura  caso
## 1  1   23    1   80    180 FALSE
## 2  2   45    0   60    160 FALSE
## 3  2   67    1   70    200  TRUE
## 4  4  201    0   50    140 FALSE
## 5  4   23    1   55    300  TRUE
```

- Chequear efectividad de los filtros (si es que los hay): próximo taller
- Chequear que las variables estén en un formato correcto para el análisis: próximo taller

¿Qué otros temas podríamos validar?

Exportar una base de datos

Podemos guardar nuestra base como texto plano:

```
write.table(data, file="data.txt", sep="\t")
```

Podemos guardar en un archivo separado por comas (`csv`):

```
# Separado por comas
write.csv(data, file="data.csv", row.names = FALSE)

# Otras opciones de separación
write.csv(data, file="data.csv", row.names = FALSE, sep = "-")
```

¿Qué ventaja tienen estos formatos?

Podemos guardar en un `excel` tradicional:

```
library(openxlsx)
write.xlsx(data, file="data.xlsx")
```

Exportar una base de datos

Podemos guardar en el formato de SPSS:

```
library(haven)
write_sav(data, "data.sav")
```

Podemos guardar en el formato de SAS:

```
write_sas(data, "data.sas7bdat")
```

Podemos guardar en el formato de Stata:

```
write_dta(data, "data.dta")
```

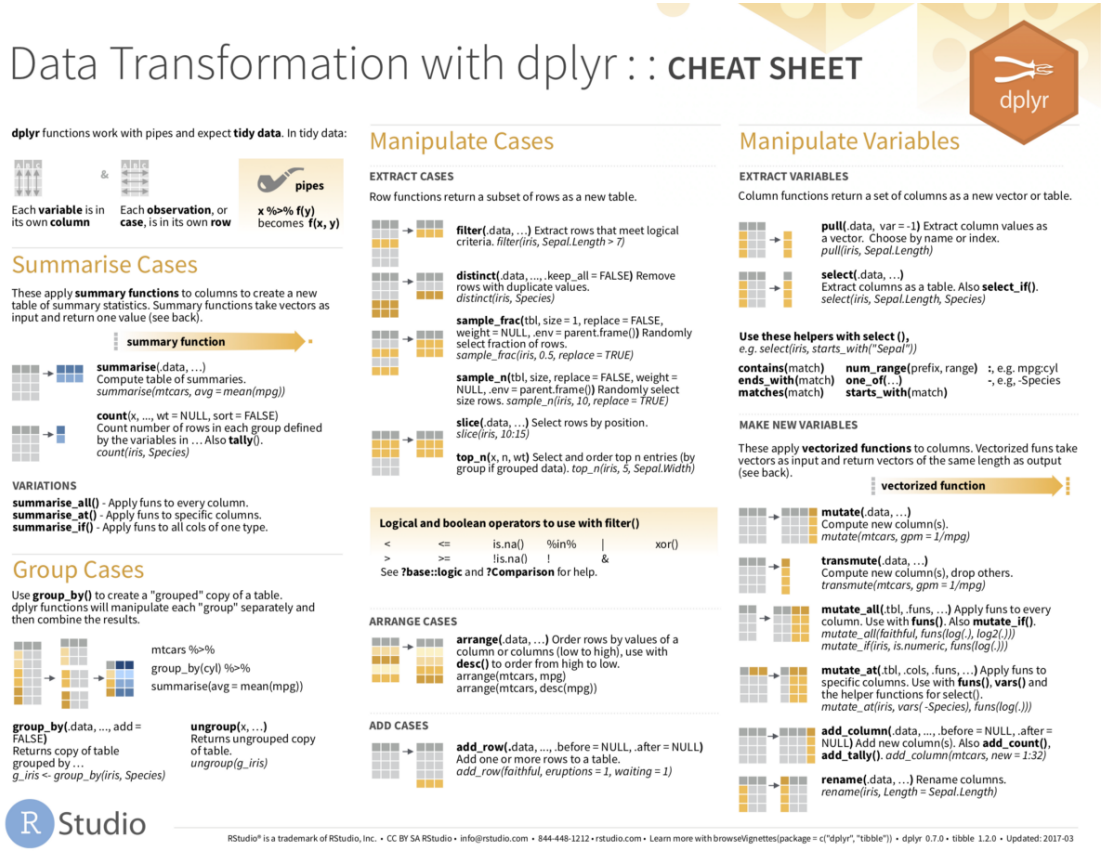
Podemos guardar en el formato de R:

```
save(data, file="data.Rdata")
```

Próximo taller

- Encadenar operaciones: `pipe %>%`
- Filtrar y ordenar datos: `filter()` & `arrange`
- Seleccionar columnas: `select()`
- Generar nuevas variables: `mutate` & operador `$`
- Recodificar y etiquetar: `indexar code` & `dplyr`)

Próximo taller



Data Transformation with dplyr : : **CHEAT SHEET**

dplyr

dplyr functions work with pipes and expect **tidy data**. In tidy data:

Each **variable** is in its own **column** & Each **observation**, or **case**, is in its own **row**

pipes
x %>% f(y) becomes **f(x, y)**

Summarise Cases

These apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

summary function

summarise(.data, …)
Compute table of summaries.
summarise(mtcars, avg = mean(mpg))

count(x, ..., wt = NULL, sort = FALSE)
Count number of rows in each group defined by the variables in … Also **tally**().
count(iris, Species)

VARIATIONS

summarise_all() - Apply funs to every column.
summarise_at() - Apply funs to specific columns.
summarise_if() - Apply funs to all cols of one type.

Group Cases

Use **group_by()** to create a "grouped" copy of a table. dplyr functions will manipulate each "group" separately and then combine the results.

mtcars %>%
group_by(cyl) %>%
summarise(avg = mean(mpg))

group_by(.data, ..., add = FALSE)
Returns copy of table grouped by …
g_iris <- group_by(iris, Species)

ungroup(x, …)
Returns ungrouped copy of table.
ungroup(g_iris)

Manipulate Cases

EXTRACT CASES

Row functions return a subset of rows as a new table.

filter(.data, …) Extract rows that meet logical criteria. *filter(iris, Sepal.Length > 7)*

distinct(.data, ..., .keep_all = FALSE) Remove rows with duplicate values.
distinct(iris, Species)

sample_frac(tbl, size = 1, replace = FALSE, weight = NULL, .env = parent.frame()) Randomly select fraction of rows.
sample_frac(iris, 0.5, replace = TRUE)

sample_n(tbl, size, replace = FALSE, weight = NULL, .env = parent.frame()) Randomly select size rows. *sample_n(iris, 10, replace = TRUE)*

slice(.data, …) Select rows by position.
slice(iris, 10:15)

top_n(x, n, wt) Select and order top n entries (by group if grouped data). *top_n(iris, 5, Sepal.Width)*

Logical and boolean operators to use with filter()

<	<=	is.na()	%in%	|	xor()
>	>=	!is.na()	!	&	

See **?base::logic** and **?Comparison** for help.

ARRANGE CASES

arrange(.data, …) Order rows by values of a column or columns (low to high), use with **desc()** to order from high to low.
arrange(mtcars, mpg)
arrange(mtcars, desc(mpg))

ADD CASES

add_row(.data, ..., .before = NULL, .after = NULL) Add one or more rows to a table.
add_row(faithful, eruptions = 1, waiting = 1)

Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

pull(.data, var = -1) Extract column values as a vector. Choose by name or index.
pull(iris, Sepal.Length)

select(.data, …)
Extract columns as a table. Also **select_if()**.
select(iris, Sepal.Length, Species)

Use these helpers with select (),
e.g. select(iris, starts_with("Sepal"))

contains(match)
ends_with(match)
matches(match)
num_range(prefix, range)
one_of(…)
starts_with(match)
:, e.g. mpg:cyl
-, e.g, -Species

MAKE NEW VARIABLES

These apply **vectorized functions** to columns. Vectorized funs take vectors as input and return vectors of the same length as output (see back).

vectorized function

mutate(.data, …)
Compute new column(s).
mutate(mtcars, gpm = 1/mpg)

transmute(.data, …)
Compute new column(s), drop others.
transmute(mtcars, gpm = 1/mpg)

mutate_all(.tbl, .funs, …) Apply funs to every column. Use with **funs()**. Also **mutate_if()**.
mutate_all(faithful, funs(log(.), log2(.)))
mutate_if(iris, is.numeric, funs(log(.)))

mutate_at(.tbl, .cols, .funs, …) Apply funs to specific columns. Use with **funs()**, **vars()** and the helper functions for select().
mutate_at(iris, vars(-Species), funs(log(.)))

add_column(.data, ..., .before = NULL, .after = NULL) Add new column(s). Also **add_count()**, **add_tally()**. *add_column(mtcars, new = 1:32)*

rename(.data, …) Rename columns.
rename(iris, Length = Sepal.Length)

RStudio

Pueden adelantar revisando el siguiente enlace

Donación

Se donará el 25% de lo recaudado en el taller: $350.000.

Se está gestionando cajas de alimentos para personas con problemas económicos en La comuna de la Pintana, sector el castillo.

¡Se les avisará cuando se entreguen estas cajas! (en las próximas 2 semanas)

Sugerencias: ✉ Buzón anónimo

¡Gracias!

Presentación generada en:

jdconejeros@uc.cl joseconejerosp

JDConejeros Jose_Conejeros