

COMPUTACIÓN ESTADÍSTICA

EPG3308

06 IMPORTACIÓN DE DATOS **readr readxl httr dbplyr**

JOSHUA KUNST
@JBKUNST

2023-04-04



IMPORTACIÓN DE DATOS

ORÍGENES DE INFORMACIÓN

Los datos pueden venir de muchas fuentes:

- Archivo de texto. <https://raw.githubusercontent.com/jbkunst/usach-ingemat-intro-elementos-ds-202201/main/data/pollution.csv>
- Planillas, archivos excel.
- SPSS (?).
- Bases de Datos. Esto es todo un mundo, existen muchos motores de bases de datos.
- Desde una página web: <https://www.reclamos.cl/empresa/uber-eats>
- Una API:
<https://climatologia.meteochile.gob.cl/application/productos/datosRecientesEma/330020/2022/05>
- Algo más?

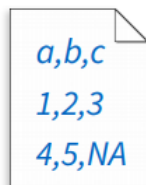
paquete::funcion PARA CADA TIPO DE ORIGEN

Origen	paquete::funcion
xlsx	readxl::read_excel
txt	readr::read_delim / data.table::fread
csv	readr::read_csv / data.table::fread
Bases de datos	Paquete dbplyr
Archivo SPSSsav	haven::read_sav
API	httr::GET
Página web	rvest::read_html

En general Bases de datos es más general dado que existen distintos motores de bases de datos.

Más info en <https://dbplyr.tidyverse.org/articles/dbplyr.html>

read_csv / read_delim



a,b,c
1,2,3
4,5,NA



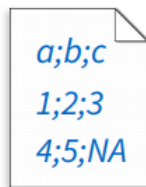
A	B	C
1	2	3
4	5	NA

Comma Delimited Files

read_csv("file.csv")

To make file.csv run:

```
write_file(x = "a,b,c\n1,2,3\n4,5,NA", path = "file.csv")
```



a;b;c
1;2;3
4;5;NA

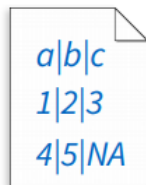


A	B	C
1	2	3
4	5	NA

Semi-colon Delimited Files

read_csv2("file2.csv")

```
write_file(x = "a;b;c\n1;2;3\n4;5;NA", path = "file2.csv")
```



a|b|c
1|2|3
4|5|NA

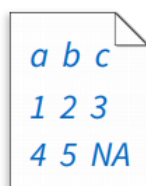


A	B	C
1	2	3
4	5	NA

Files with Any Delimiter

read_delim("file.txt", delim = "|")

```
write_file(x = "a|b|c\n1|2|3\n4|5|NA", path = "file.txt")
```



a b c
1 2 3
4 5 NA



A	B	C
1	2	3
4	5	NA

Fixed Width Files

read_fwf("file.fwf", col_positions = c(1, 3, 5))

```
write_file(x = "a b c\n1 2 3\n4 5 NA", path = "file.fwf")
```

Tab Delimited Files

read_tsv("file.tsv") Also **read_table()**.

```
write_file(x = "a\tb\tc\n1\t2\t3\n4\t5\tNA", path = "file.tsv")
```

BASES DE DATOS **dbplyr**

```
library(dplyr)
library(dbplyr)
library(DBI)
library(RMariaDB)

con <- dbConnect(
  RMariaDB::MariaDB(),
  host = "mysql-rfam-public.ebi.ac.uk",
  db = "Rfam",
  user = "rfamro",
  port = 4497
)

tbl(con, "taxonomy") |>
  select(ncbi_id, species, tax_string) |>
  filter(ncbi_id == 10116) |>
  show_query() |>
  collect()
```

```
## <SQL>
## SELECT `ncbi_id`, `species`, `tax_string`
## FROM `taxonomy`
## WHERE (`ncbi_id` = 10116.0)

## # A tibble: 1 × 3
##   ncbi_id species                tax_string
##   <int> <chr>                  <chr>
## 1    10116 Rattus norvegicus (Norway rat) Eukaryota;
```

HTML `rvest::read_html`

```
library(rvest)
library(lubridate)

url <- "https://www.reclamos.cl/empresa/pontific:

read_html(url) |>
  html_table() |>
  first() |>
  set_names(c("fecha", "reclamo", "cantidad")) |>
  mutate(fecha = dmy(fecha)) |>
  arrange(fecha, cantidad) |>
  ggplot(aes(fecha, cantidad)) +
  geom_line(color = "gray80", size = 1.2) +
  geom_smooth(alpha = 0.25, size = 1.2) +
  coord_cartesian(ylim = c(0, NA)) +
  theme_minimal()
```

