Ciencia de datos en R

Big Data: Marco conceptual, técnicas y aplicaciones

Clase 2

Universidad de San Andrés

Temario de hoy

- ¿Qué son los paquetes?
- Nombrar variables
- Carga de datos .csv o .xlsx
- Verbos tidyverse

¿Qué son los paquetes?

- ➔ Son conjuntos de código, datos, documentación y tests que otra persona o personas han desarrollado y que nosotros podemos usar gratuitamente.
- ➔ Es la unidad fundamental de código "*compartible*", a través de las funciones.
- ➔ Arranquemos instalando dos que vamos a usar en la próximas filminas.
- ➔ La instalación es por única vez en cada equipo, después simplemente se cargan.

```
> install.packages("tidyverse")
> install.packages("readxl")

> library(tidyverse)
> library(readxl)
```

Tidyverse

- ➔ Es un “*metapaquete*”: una colección curada de paquetes diseñados para Ciencia de Datos.
- ➔ Todos los paquetes tienen una misma filosofía, gramática y estructura de datos.

Más info: Tidyverse, Hadley Wickham & RStudio

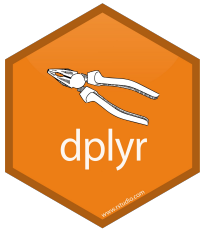
Carga de datos

➔ Vamos a trabajar con distintos tipos de archivos:

- CSV (*Comma-separated values)*
  - Altamente utilizado en el mundo de las ciencia de datos
  - Fácil de abrir por distintos softwares (menos Excel)
- XLSX (Microsoft Excel)

```
> escuelas_csv <- read.csv(“nombre_archivo.csv”)
> escuelas_xls <- read_excel(“nombre_archivo.xls”)
```

dplyr



- ➔ “Gramática” para manipular datos.
- ➔ Usa “verbos” que modifican dataframes.
  - ◆ select() seleccionar columnas.
  - ◆ filter() filtrar filas
  - ◆ group_by() agrupar los datos según los valores de las columnas.
  - ◆ summarise() reducir muchos valores a un estadístico.
  - ◆ mutate() agregar nuevas columnas.
  - ◆ arrange() reordenar las filas.
  - ◆ rename() cambiar de nombre algunas variables.

%>% (pipe)

➔ Aplica una función al resultado de la función anterior.


```
torta_1_2_3_4 <- crema_un_pote %>%
                    agregar(huevos, cantidad = "4") %>%
                    agregar(esencia de vainilla, cantidad = "1 cucharada") %>%
                    agregar(azucar, cantidad = "2 potes") %>%
                    agregar(harina, cantidad = "3 potes") %>%
                    batir() %>%
                    enmantecar_fuente() %>%
                    verter_mezcla_fuente() %>%
                    hornear(tiempo = "45 minutos", temperatura = "150 grados") %>%
                    disfrutar()
```

Más info: Torta 1234, quierocakes.com

Nombrar variables

“There are only two hard things in Computer Science: cache invalidation and naming things.”
— Phil Karlton

Guía de estilo para nombrar variables:

- ➔ Usar **sustantivos** para las variables y **verbos** para las funciones.
- ➔ Usar nombres que sean **concisos y significativos!**!
- ➔ Usar solo letras y números en **minúsculas**
- ➔ Separar las palabras por **“_”**. Ej: *escuelas_en_pba*.

Más info: The tidyverse style guide, Hadley Wickham

Comparaciones

- ➔ == igual a
- ➔ != no igual a
- ➔ > mayor a
- ➔ >= mayor o igual a
- ➔ < menor a
- ➔ <= menor o igual a

Operadores lógicos

- ➔ a & b a y b
- ➔ a | b a ó b
- ➔ a & !b a, y no b
- ➔ !a & b no a, y b
- ➔ !(a & b) no (a y b)