



Week 11

R

Day 3

DPLYR & GGLOT2

Tidyverse

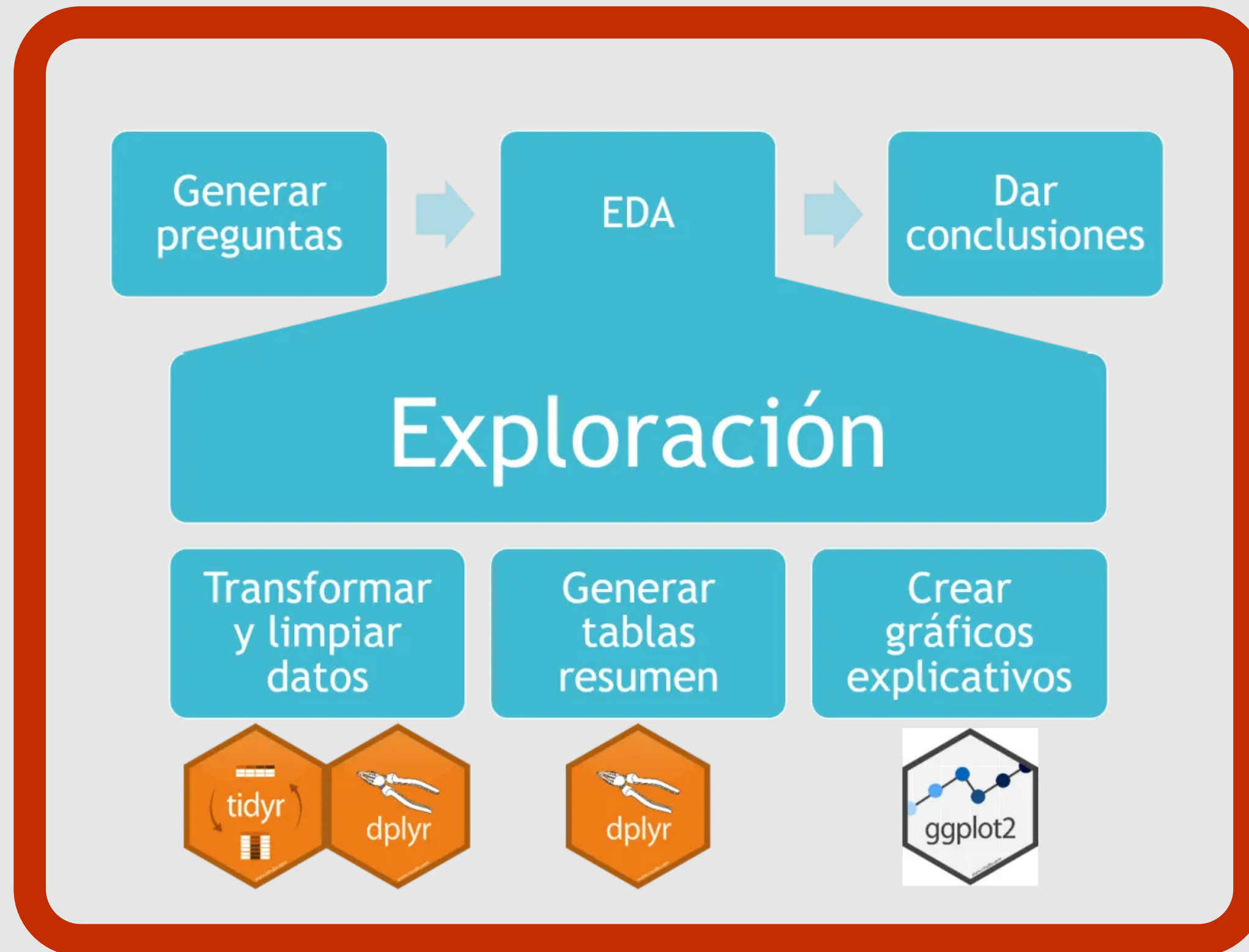
<https://www.tidyverse.org/>



Es una colección de paquetes de R diseñados para Data Science.

Todos los paquetes utilizan la misma filosofía de diseño, gramática y estructura de datos.

Tidyverse



dplyr

<https://dplyr.tidyverse.org/>

Ayuda con problemas comunes de manipulación de datos en un lenguaje basado en acciones sobre los datos.

Estas acciones se conocen como verbos.

- **mutate()** crea nuevas variables
- **select()** toma variables según nombre
- **filter()** toma variables según condición
- **summarise()** reduce muchos valores en un único resumen
- **arrange()** cambia el orden de los datos
- **group_by()** operaciones por grupos

El operador pipe %>%

nos permite escribir una secuencia de

operaciones de izquierda a derecha:

```
gapminder %>%  
  filter(year == 2007) %>%  
  mutate(lifeExpMonths = 12 * lifeExp) %>%  
  arrange(desc(lifeExpMonths)) %>%  
  select(continent)
```

ggplot2

<https://ggplot2.tidyverse.org/>

Paquete de visualización.

Utiliza la gramática de los gráficos.

Función para graficar

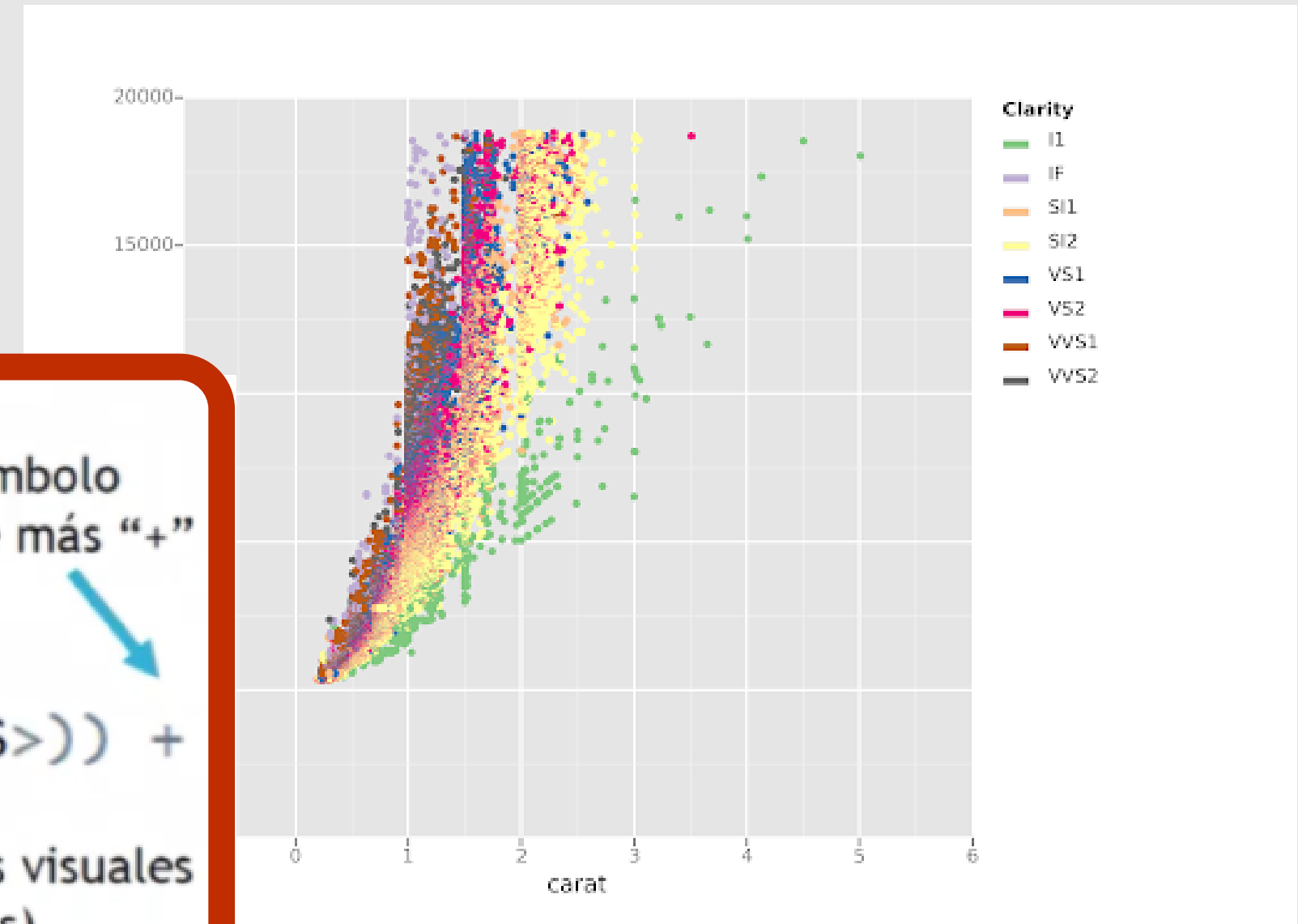
datos a graficar

Símbolo de más “+”

```
ggplot(data = <DATA>,  
       mapping = aes(<MAPPINGS>)) +  
  <GEOM_FUNCTION>()
```

Geometría (tipo de gráfica)

Elementos visuales (aesthetics)



readr

<https://readr.tidyverse.org/>




Permite leer de manera amigable y rápida archivos de texto plano (csv)

- **read_csv** para importar un archivo .CSV
- **write_csv** para exportar un archivo

Ambas funciones son 10 veces más rápidas que las versiones de R base.

tidyr

<https://tidyr.tidyverse.org/>



Tiene como finalidad el tomar datos no tidy y transformarlos en datos limpios y ordenados.

- **Cada columna es una variable**
- **Cada fila una observación**
- **Cada celda un valor**

Para este fin utilizaremos principalmente las funciones: **gather** y **spread**.

otros paquetes

purrr - facilita el trabajo con vectores y funciones en un lenguaje consistente

<https://purrr.tidyverse.org/>

tibble - es una reinención del dataframe. Haciendo más eficiente algunas rutinas

<https://tibble.tidyverse.org/>

stringr - paquete para trabajar con análisis de textos y manipulación de strings

<https://stringr.tidyverse.org/>

forcats - es un paquete especial para lidiar con factores y datos categóricos

<https://forcats.tidyverse.org/>

otros paquetes

(no en la versión principal de tidyverse)

para leer datos:

readxl (Excel)

haven (SPSS, Stata, SAS)

manipulación de datos:

lubridate (fechas y tiempos)

hms (para horas, minutos, segundos)

blob (para datos binarios)



You got this!



Relax & keep coding

Ejercicio práctico

Crea un archivo .R en tu proyecto de RStudio y envía el enlace de github al archivo por email.

01.

Elige alguno de los built-in datasets de R, ¡el que quieras! y realiza un análisis de este.

Requisitos mínimos.

- Aplica los seis verbos de dplyr, al menos una vez. Pero pueden ser muchas.
- Haz un pipeline de verbos usando %>% con al menos 3 de ellos.
- Haz un pipeline de verbos en el que incluyas al menos un verbo de dplyr y una función de ggplot.
- Haz al menos tres gráficas con ggplot que cuenten algo de tus datos.

02.

Define una función que coja como argumento el path donde está `sheeps.csv` y devuelva una gráfica hecha con ggplot2.

Añade un parámetro que por defecto sea False, pero que si es True aplique algún filtro a los datos gráfico diferente.

Es posible que para poder llevar a cabo la gráfica tengas que limpiar los datos y aplicar alguna función de dplyr o tidyr.

03.

Importa el archivo: `president.csv` utilizando la librería readr.

Utiliza al menos una función de la librería tidyr para hacer estos datos tidyer.