INSP/ESPM
Escuela de Salud Pública de México

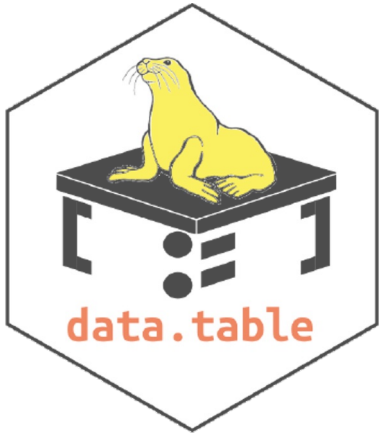
Introducción al Tidy Data

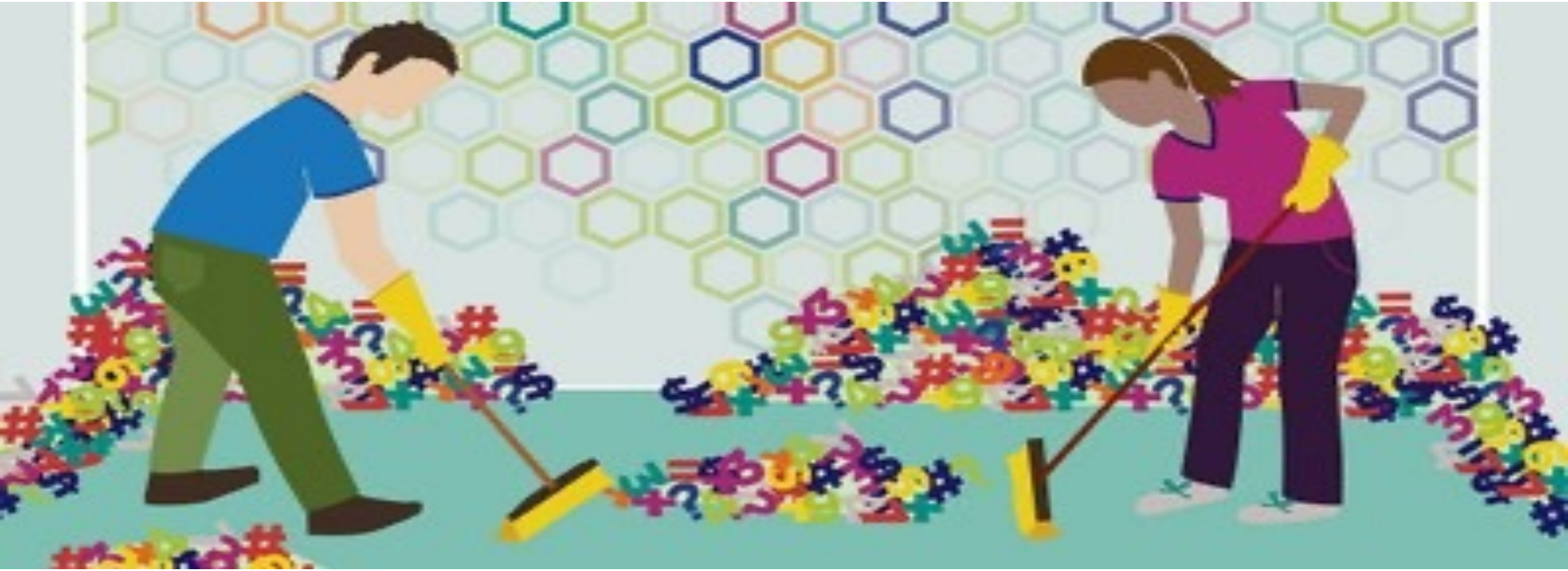
José Luis Texcalac Sangrador

Procesamiento y visualización de datos espaciales en R

Tidyverse

- Actualmente en **R** predominan tres entornos de trabajo, con sus respectivas librerías, que nos permiten programar y manipular la información.
- Cada una tiene sus propias ventajas y desventajas, así como la forma de abordar la manipulación de datos.
- Para este curso, adoptaremos la filosofía de Tidyverse.

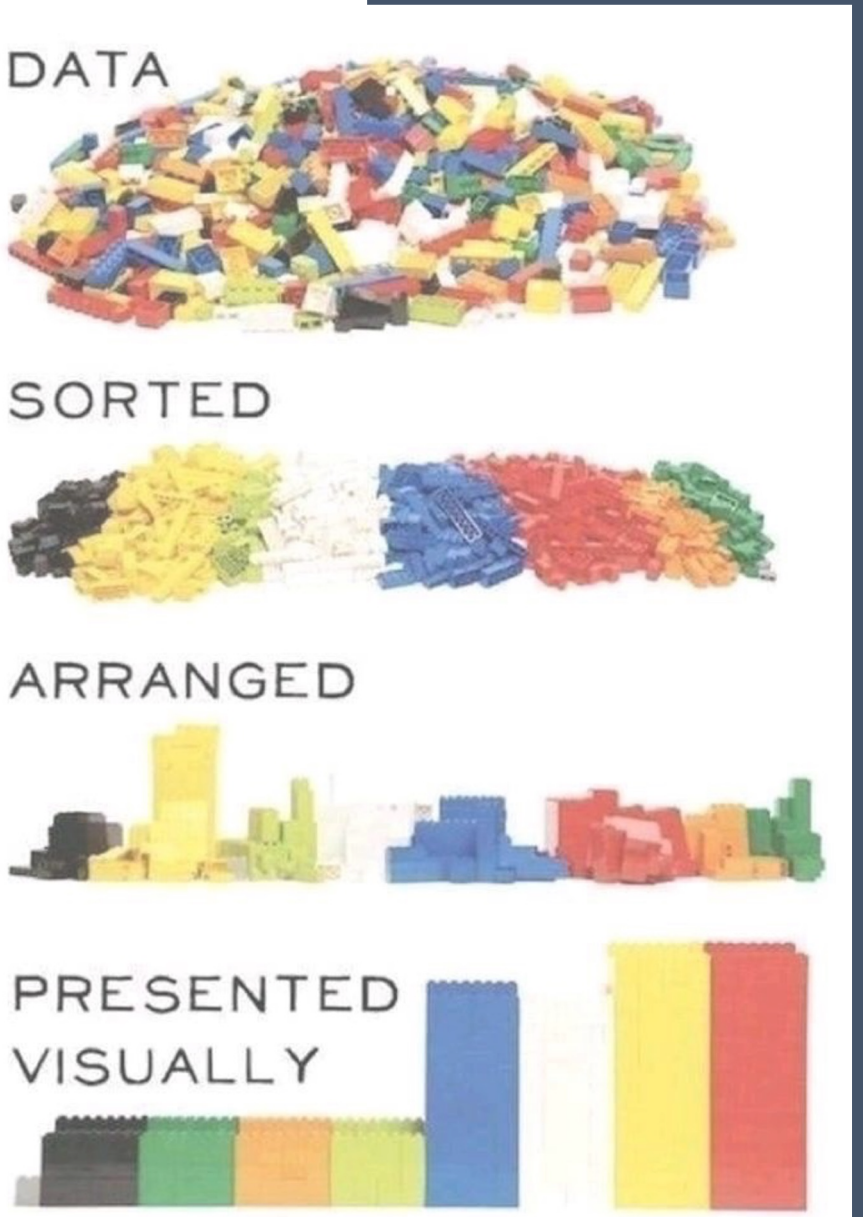




Hadley Wickham, gestor de esta filosofía en la ciencia de datos, recuerda que uno de sus principales problemas en el trabajo con datos era la gran cantidad de tiempo que se tenía que invertir para ordenar y analizar la información través de las herramientas de R y RStudio.

Tidy data

- Este antecedente le llevó a gestar la idea de data tidying o tidy data, es decir:
- "Estructurar un conjunto de datos (data) para facilitar su análisis"
- El concepto de tidy data implica una forma estandarizada de vincular la estructura de un conjunto de datos (su disposición física), con su semántica (su significado).



tidy data

Cada columna es una variable

nombre	edad	peso
Kevin	17	64.3
Brayan	16	61.7
Kimberly	15	51.9
Britany	16	59.3
Brandon	17	69.1
Melany	16	61.6

Cada celda es un valor e indica la intersección entre una variable y una observación

Cada fila es una observación

tidy data

Dataset: Colección de valores (tabla, malla de datos, data frame)

nombre	edad	peso
Kevin	17	64.3
Brayan	16	61.7
Kimberly	15	51.9
Britany	16	59.3
Brandon	17	69.1
Melany	16	61.6

tidy data

Dataset: Colección de valores (tabla, malla de datos, data frame)

Variable: Contiene todos los valores que miden el mismo atributo entre todas las unidades de observación

nombre	edad	peso
Kevin	17	64.3
Brayan	16	61.7
Kimberly	15	51.9
Britany	16	59.3
Brandon	17	69.1
Melany	16	61.6

tidy data

Dataset: Colección de valores (tabla, malla de datos, data frame)

Variable: Contiene todos los valores que miden el mismo atributo entre todas las unidades de observación

Observación: Contiene todos los valores medidos, en todos los atributos, en la misma unidad de observación

nombre	edad	peso
Kevin	17	64.3
Brayan	16	61.7
Kimberly	15	51.9
Britany	16	59.3
Brandon	17	69.1
Melany	16	61.6

tidy data

Dataset: Colección de valores (tabla, malla de datos, data frame)

Variable: Contiene todos los valores que miden el mismo atributo entre todas las unidades de observación

Observación: Contiene todos los valores medidos, en todos los atributos, en la misma unidad de observación

Valor: Intersección entre la variable y la observación

nombre	edad	peso
Kevin	17	64.3
Brayan	16	61.7
Kimberly	15	51.9
Britany	16	59.3
Brandon	17	69.1
Melany	16	61.6

messy data

Los encabezados son valores, no atributos

nombre	edad	peso	lacio	rizado	ondulado
Kevin	17	64.3	0	1	0
Brayan	16	61.7	0	0	0
Kimberly	15	51.9	1	0	0
Britany	16	59.3	0	0	0
Brandon	17	69.1	0	0	1
Melany	16	61.6	0	0	0

INSP/ESPM
Escuela de Salud Pública de México

Datos wide y long

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

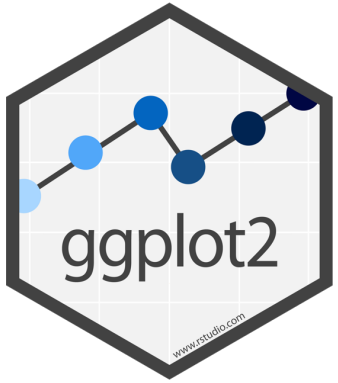
Datos wide y long

Site	2013	2014	2015
CAM	51.0	42.8	39.9
FAC	48.3	39.0	36.6
IZT	44.6	39.3	35.0

Site	Year	PM10
CAM	2013	51.0
FAC	2013	48.3
IZT	2013	44.6
CAM	2014	42.8
FAC	2014	39.0
IZT	2014	39.3
CAM	2015	39.9
FAC	2015	36.6
IZT	2015	35.0

INSP/ESPM
Escuela de Salud Pública de México

Paquetes base



Su turno...

- Desde la consola de **R** instale los siguientes paquetes del CRAN:
  - tidyverse, lubridate, readxl, haven, janitor.

```
> install.packages("tidyverse", dependencies = TRUE)
```


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

Es equivalente a:

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("stringr")
install.packages("forcats")
```

```
library(tidyverse)
```

Es equivalente a:

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(readr)
library(purrr)
library(tibble)
library(stringr)
library(forcats)
```

INSP/ESPM
Escuela de Salud Pública de México

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

Es equivalente a:

```
install.packages("ggplot2")

install.packages("dplyr")

install.packages("tidyr")

install.packages("readr")

install.packages("purrr")

install.packages("tibble")

install.packages("stringr")

install.packages("forcats")
```

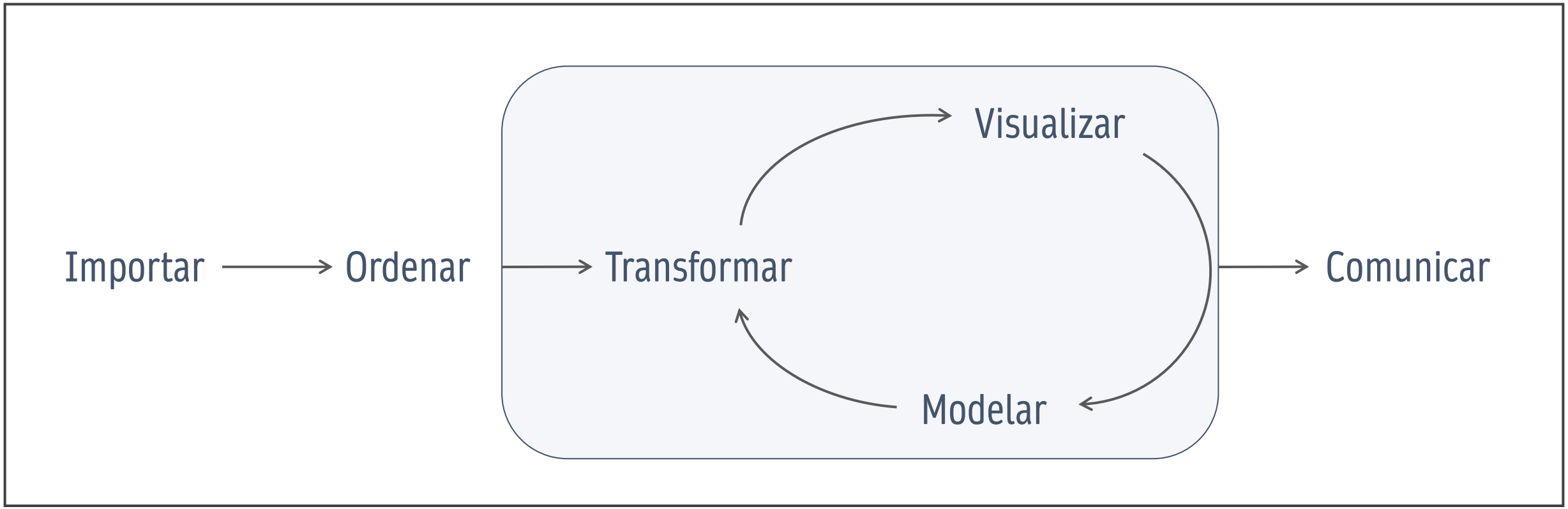
```
library(tidyverse)
```

Es equivalente a:

```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R es un software libre y viene sin GARANTIA ALGUNA.
Usted puede redistribuirlo bajo ciertas circunstancias.
Escriba 'license()' o 'licence()' para detalles de distribucion.

R es un proyecto colaborativo con muchos contribuyentes.
Escriba 'contributors()' para obtener más información y
'citation()' para saber cómo citar R o paquetes de R en publicaciones.

Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
Escriba 'q()' para salir de R.

> library(tidyverse)
── Attaching packages ─────────────────────────────────── tidyverse 1.3.1 ──
✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.4     ✓ dplyr   1.0.7
✓ tidyr   1.1.3     ✓ stringr 1.4.0
✓ readr   2.0.1     ✓ forcats 0.5.1
── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

Librerías ampliadas tidyverse

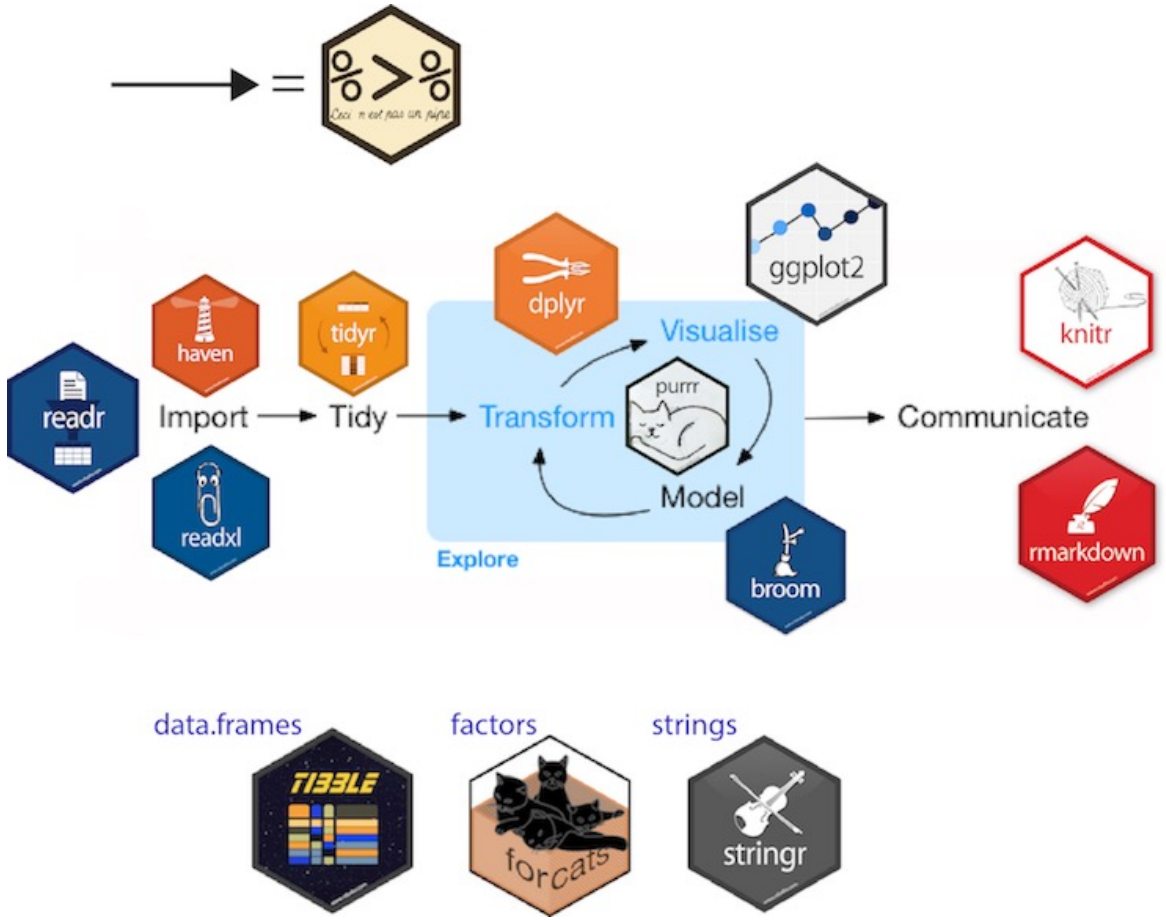
Flujo de trabajo



Programar

INSP/ESPM
Escuela de Salud Pública de México

Flujo de trabajo usando paquetes



Pipe

El operador %>% simplifica y concatena múltiples funciones

```
mi_dia <- veo_tv(paseo_perro(regreso(trabajo(traslado(despierto(🙂))))))
```

```
🙂 %>%
  despierto %>%
  traslado %>%
  trabajo %>%
  regreso %>%
  paseo_perro %>%
  veo_tv
```

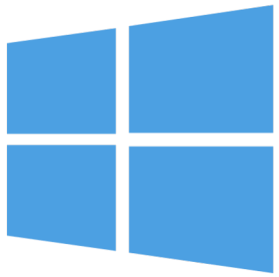
```
malla_datos %>%
  filtro %>%
  genero_variables %>%
  agrupo %>%
  paso_a_wide %>%
  genero_variables %>%
  selecciono_columnas
```

Insertar pipe

command + shift + M

Mac

Ctrl + Shift + M



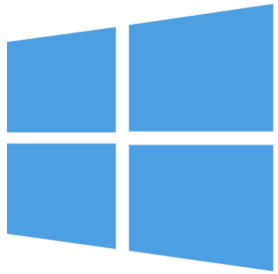
Windows

Insertar chunk

⌘ command + ⌥ alt + I

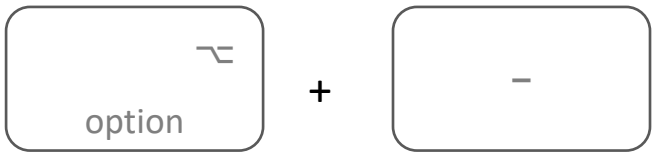
Mac

Ctrl + Alt + I

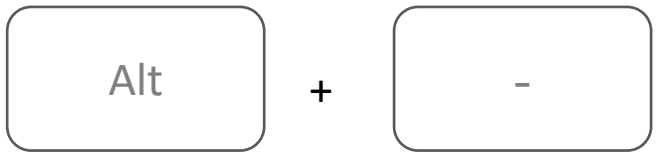


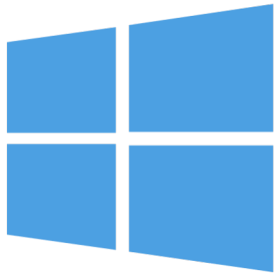
Windows

Insertar operador de asignación <-



Mac





Windows

INSP/ESPM
Escuela de Salud Pública de México

Using = instead of <- for assignment