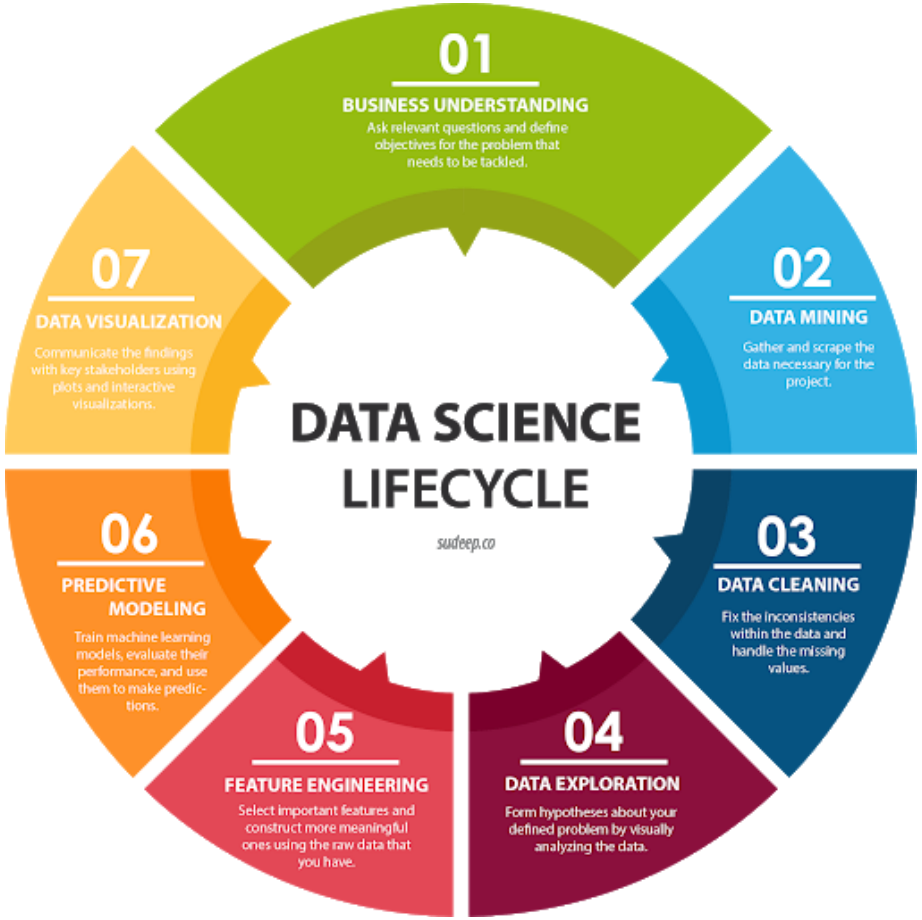
Dplyr

Manipulación de datos con R

Ciencia de Datos

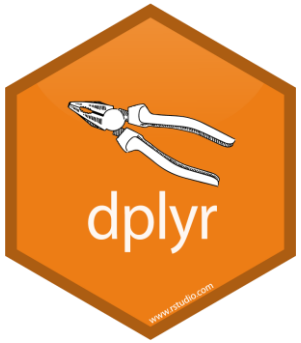


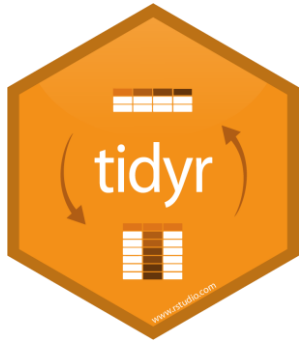
Tidyverse

- Conjuntos de librerías para hacen fácil la ciencia de datos:
  - Todas las funciones reciben como primer parametro un dataframe.

```
> nycflights13::flights
# A tibble: 336,776 x 19
    year month   day dep_t… sched_… dep_d… arr_… sched… arr_d… carr… flig… tail… orig… dest  air_…
   <int> <int> <int>  <int>   <int>  <dbl> <int>  <int>  <dbl> <chr> <int> <chr> <chr> <chr> <dbl>
 1  2013     1     1    517     515   2.00   830    819   11.0 UA     1545 N142… EWR   IAH   227
 2  2013     1     1    533     529   4.00   850    830   20.0 UA     1714 N242… LGA   IAH   227
 3  2013     1     1    542     540   2.00   923    850   33.0 AA     1141 N619… JFK   MIA   160
 4  2013     1     1    544     545  -1.00  1004   1022  -18.0 B6      725 N804… JFK   BQN   183
 5  2013     1     1    554     600  -6.00   812    837  -25.0 DL      461 N668… LGA   ATL   116
 6  2013     1     1    554     558  -4.00   740    728   12.0 UA     1696 N394… EWR   ORD   150
 7  2013     1     1    555     600  -5.00   913    854   19.0 B6      507 N516… EWR   FLL   158
 8  2013     1     1    557     600  -3.00   709    723  -14.0 EV     5708 N829… LGA   IAD    53.0
 9  2013     1     1    557     600  -3.00   838    846   -8.00 B6      79 N593… JFK   MCO   140
10  2013     1     1    558     600  -2.00   753    745    8.00 AA     301 N3AL… LGA   ORD   138
# ... with 336,766 more rows, and 4 more variables: distance <dbl>, hour <dbl>, minute <dbl>,
#   time_hour <dttm>
```

Cuáles son las librerías Core



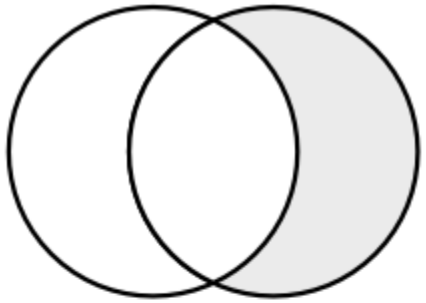


Verbos de la Manipulación de Datos

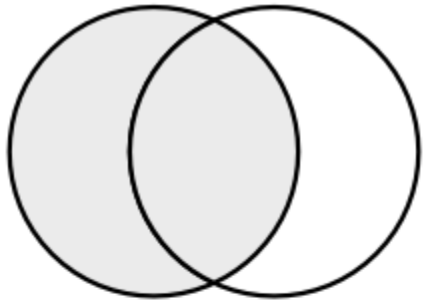
- **filter()**: Seleccionar casos (filas) dependiendo de sus valores.
- **select()** y **rename()**: Seleccionar variables basado en sus nombres.
- **mutate()** Adicionar nuevas variables que son funciones de variables existentes.
- **summarise()**: Resumir multiples valores a un único valor.
- **arrange()**: Ordenamiento de los casos (filas).

Todos los verbos, sobretodo mutate, summarize y arrange interactúan con

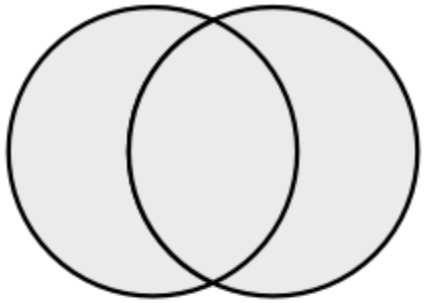
Operadores Filter



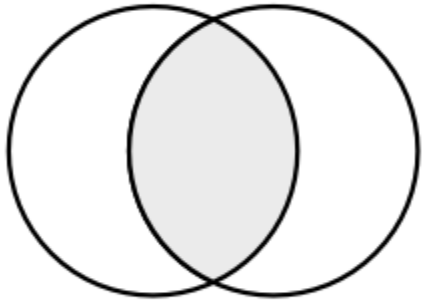
y & !x



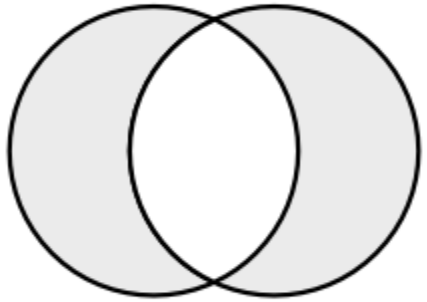
x



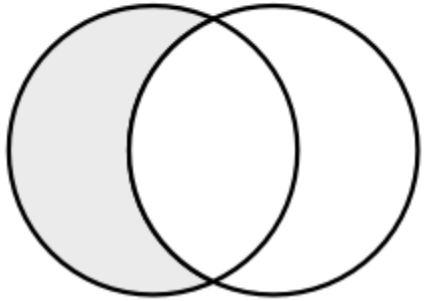
x | y



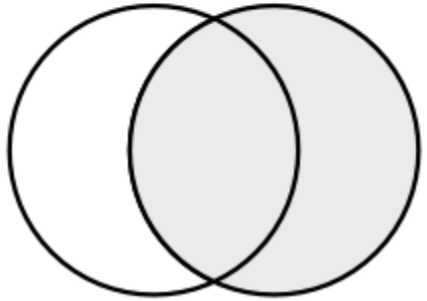
x & y



xor(x, y)

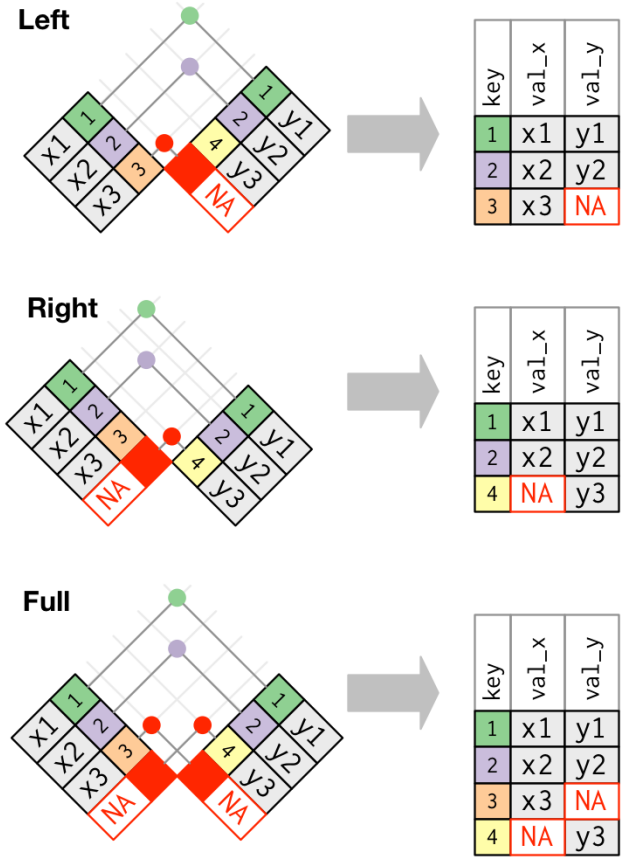


x & !y



y

Joins Dplyr


Left
x1
x2
x3
y1
y2
y3
NA
key
val_x
val_y
1 x1 y1
2 x2 y2
3 x3 NA
Right
x1
x2
x3
NA
y1
y2
y3
key
val_x
val_y
1 x1 y1
2 x2 y2
4 NA y3
Full
x1
x2
x3
NA
NA
y1
y2
y3
key
val_x
val_y
1 x1 y1
2 x2 y2
3 x3 NA
4 NA y3

TidyR

gather() and spread()

	Year	Month	Day	Element	Temp
1	2015	1	1	tmax	78
2	2015	1	1	tmin	72
3	2015	2	2	tmax	82
4	2015	2	2	tmin	74
5	2015	4	4	tmax	81
6	2015	4	4	tmin	71
7	2015	6	3	tmax	80
8	2015	6	3	tmin	71

spread(myData, Element, Temp)

	Year	Month	Day	tmax	tmin
1	2015	1	1	78	72
2	2015	2	2	82	74
3	2015	4	4	81	71
4	2015	6	3	80	71

```
myData %>% gather("Element","Temp",-Year,-Month,-Day) %>%
           rename(c(variable = "Element",value="Temp"))
```