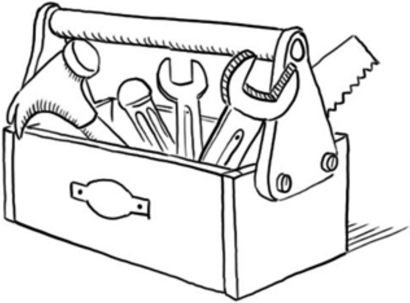
# Caja de Herramientas: R@FSOC

TAO

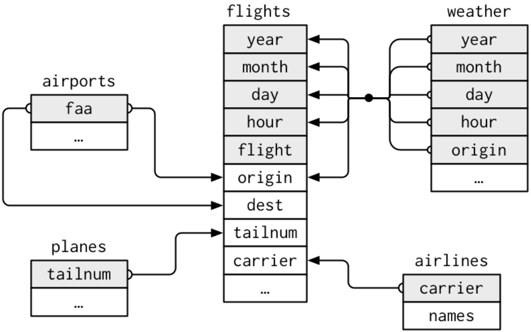
28/05/2021

# Caja de Herramientas



# Uniendo Tablas

# Uniendo Tablas



# Librerias

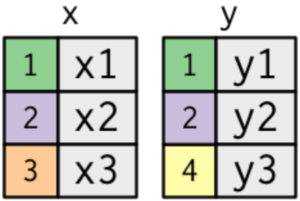
```
## -- Attaching packages ---------------------------------

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

# Explicado los Joins

# Ejemplo



x

| | |
|---|---|
| 1 | x1 |
| 2 | x2 |
| 3 | x3 |

y

| | |
|---|---|
| 1 | y1 |
| 2 | y2 |
| 4 | y3 |

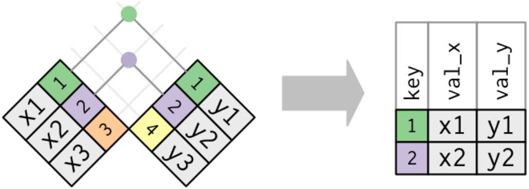
```
x <- tribble(
  ~key, ~val_x,
     1, "x1",
     2, "x2",
     3, "x3"
)
```

```
y <- tribble(
  ~key, ~val_y,
     1, "y1",
     2, "y2",
     4, "y3"
)
```

# Inner join

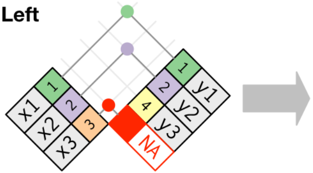
```
x %>%
  inner_join(y, by = "key")
```

```
## # A tibble: 2 x 3
##     key val_x val_y
##   <dbl> <chr> <chr>
## 1     1 x1    y1
## 2     2 x2    y2
```



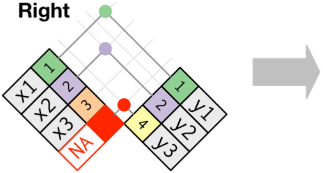
| key | val_x | val_y |
|---|---|---|
| 1 | x1 | y1 |
| 2 | x2 | y2 |

# Otros join



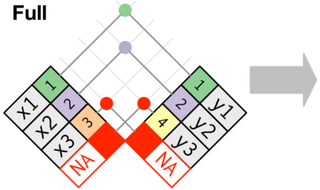
Left

| key | val_x | val_y |
|---|---|---|
| 1 | x1 | y1 |
| 2 | x2 | y2 |
| 3 | x3 | NA |



Right

| key | val_x | val_y |
|---|---|---|
| 1 | x1 | y1 |
| 2 | x2 | y2 |
| 4 | NA | y3 |



Full

| key | val_x | val_y |
|---|---|---|
| 1 | x1 | y1 |
| 2 | x2 | y2 |
| 3 | x3 | NA |
| 4 | NA | y3 |

# Left join

```
x %>%
  left_join(y, by = "key")
```

```
## # A tibble: 3 x 3
##     key val_x val_y
##   <dbl> <chr> <chr>
## 1     1 x1    y1
## 2     2 x2    y2
## 3     3 x3    <NA>
```

# Right join

```r
x %>%
  right_join(y, by = "key")
```

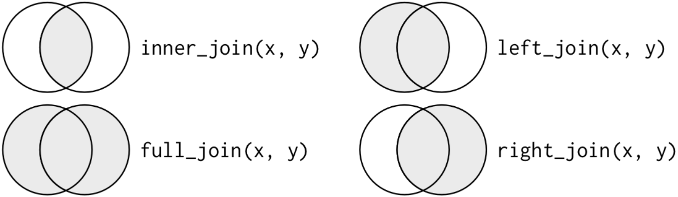
```
## # A tibble: 3 x 3
##     key val_x val_y
##   <dbl> <chr> <chr>
## 1     1 x1    y1
## 2     2 x2    y2
## 3     4 <NA>  y3
```

# Full join

```
x %>%
  full_join(y, by = "key")
```

```
## # A tibble: 4 x 3
##     key val_x val_y
##   <dbl> <chr> <chr>
## 1     1 x1    y1
## 2     2 x2    y2
## 3     3 x3    <NA>
## 4     4 <NA>  y3
```

# Resumen

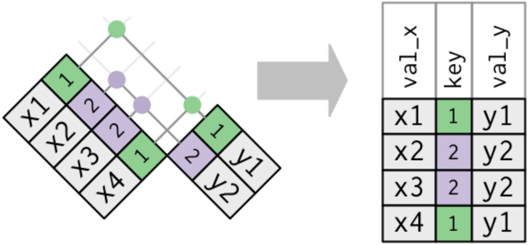


# Duplicados: Caso 1

```
x <- tribble(
  ~key, ~val_x,
     1, "x1",
     2, "x2",
     2, "x3",
     1, "x4"
)
```

```
y <- tribble(
  ~key, ~val_y,
     1, "y1",
     2, "y2"
)
```

# Duplicados: Caso 1



| val_x | key | val_y |
|---|---|---|
| x1 | 1 | y1 |
| x2 | 2 | y2 |
| x3 | 2 | y2 |
| x4 | 1 | y1 |

```r
inner_join(x, y, by = "key")
```

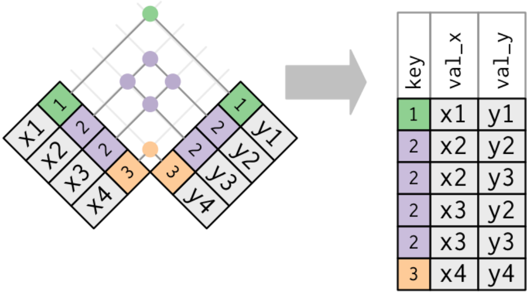
```
## # A tibble: 4 x 3
##     key val_x val_y
##   <dbl> <chr> <chr>
## 1     1 x1    y1
## 2     2 x2    y2
## 3     2 x3    y2
## 4     1 x4    y1
```

# Duplicados: Caso 2

```r
x <- tribble(
  ~key, ~val_x,
     1, "x1",
     2, "x2",
     2, "x3",
     3, "x4"
)
```

```r
y <- tribble(
  ~key, ~val_y,
     1, "y1",
     2, "y2",
     2, "y3",
     3, "y4"
)
```

# Duplicados: Caso 2



| key | val_x | val_y |
|---|---|---|
| 1 | x1 | y1 |
| 2 | x2 | y2 |
| 2 | x2 | y3 |
| 2 | x3 | y2 |
| 2 | x3 | y3 |
| 3 | x4 | y4 |

```r
inner_join(x, y, by = "key")
```

```
## # A tibble: 6 x 3
##     key val_x val_y
##   <dbl> <chr> <chr>
## 1     1 x1    y1
## 2     2 x2    y2
## 3     2 x2    y3
## 4     2 x3    y2
## 5     2 x3    y3
## 6     3 x4    y4
```

# Ejemplo

# Librarias

# Tablas

```
head(airlines)
```

```
## # A tibble: 6 x 2
##   carrier name
##   <chr>   <chr>
## 1 9E      Endeavor Air Inc.
## 2 AA      American Airlines Inc.
## 3 AS      Alaska Airlines Inc.
## 4 B6      JetBlue Airways
## 5 DL      Delta Air Lines Inc.
## 6 EV      ExpressJet Airlines Inc.
```

```
head(airports)
```

```
## # A tibble: 6 x 8
##   faa   name                             lat   lon   alt    tz dst   tzone
##   <chr> <chr>                          <dbl> <dbl> <dbl> <dbl> <chr> <chr>
## 1 04G   Lansdowne Airport               41.1 -80.6  1044    -5 A     America/New_Y~
## 2 06A   Moton Field Municipal Airp~  32.5 -85.7   264    -6 A     America/Chica~
## 3 06C   Schaumburg Regional             42.0 -88.1   801    -6 A     America/Chica~
## 4 06N   Randall Airport                 41.4 -74.4   523    -5 A     America/New_Y~
## 5 09J   Jekyll Island Airport           31.1 -81.4    11    -5 A     America/New_Y~
## 6 0A9   Elizabethton Municipal Air~  36.4 -82.2  1593    -5 A     America/New_Y~
```

# Tablas

```
head(planes)
```

```
## # A tibble: 6 x 9
##   tailnum  year type           manufacturer   model  engines seats speed engine
##   <chr>   <int> <chr>          <chr>          <chr>    <int> <int> <int> <chr>
## 1 N10156   2004 Fixed wing mu~ EMBRAER        EMB-1~       2    55    NA Turbo-~
## 2 N102UW   1998 Fixed wing mu~ AIRBUS INDUST~ A320-~       2   182    NA Turbo-~
## 3 N103US   1999 Fixed wing mu~ AIRBUS INDUST~ A320-~       2   182    NA Turbo-~
## 4 N104UW   1999 Fixed wing mu~ AIRBUS INDUST~ A320-~       2   182    NA Turbo-~
## 5 N10575   2002 Fixed wing mu~ EMBRAER        EMB-1~       2    55    NA Turbo-~
## 6 N105UW   1999 Fixed wing mu~ AIRBUS INDUST~ A320-~       2   182    NA Turbo-~
```

```
head(weather)
```

```
## # A tibble: 6 x 15
##   origin  year month   day  hour  temp  dewp humid wind_dir wind_speed wind_gust
##   <chr>  <int> <int> <int> <int> <dbl> <dbl> <dbl>    <dbl>      <dbl>     <dbl>
## 1 EWR     2013     1     1     1  39.0  26.1  59.4      270      10.4         NA
## 2 EWR     2013     1     1     2  39.0  27.0  61.6      250       8.06        NA
## 3 EWR     2013     1     1     3  39.0  28.0  64.4      240      11.5         NA
## 4 EWR     2013     1     1     4  39.9  28.0  62.2      250      12.7         NA
## 5 EWR     2013     1     1     5  39.0  28.0  64.4      260      12.7         NA
## 6 EWR     2013     1     1     6  37.9  28.0  67.2      240      11.5         NA
## # ... with 4 more variables: precip <dbl>, pressure <dbl>, visib <dbl>,
## #   time_hour <dttm>
```

# Tablas

```
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # ... with 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

# Duplicados

```
# Chequeo Duplicados, para entender la relación
mean(duplicated(airlines$carrier))
```

```
## [1] 0
```

```
mean(duplicated(flights$carrier))
```

```
## [1] 0.9999525
```

# Keys

```
# Chequeo las keys para union
key_1 = unique(flights$carrier)
key_2 = unique(airlines$carrier)

mean(key_1 %in% key_2)
```

```
## [1] 1
```

```
mean(key_2 %in% key_1)
```

```
## [1] 1
```

# Join

# Join

```
# Joins
Eg_1 = flights %>%
  left_join(airlines)
```

```
## Joining, by = "carrier"
```

```
# Chequeo Final
head(Eg_1)        # Chequeo de que hubo union
```

```
## # A tibble: 6 x 20
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # ... with 12 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>, name <chr>
```

```
nrow(flights)     # Chequeo cantidad de filas de flights
```

```
## [1] 336776
```

```
nrow(airlines)    # Chequeo  cantidad de filas de  airlines
```

```
## [1] 16
```

```
nrow(Eg_1)        # Chequeo  cantidad de filas de Eg_1
```

```
## [1] 336776
```

```
# airlines_sub <- airlines %>%
#   filter(carrier %in% c("AA","AS" ))
#
# airlines_sub[2,1] <- "AZ"
```