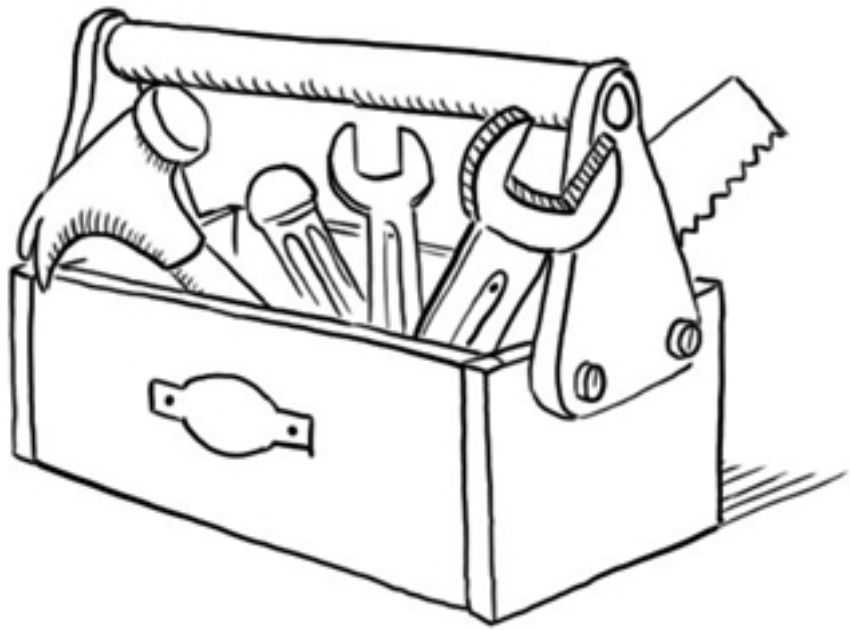
Caja de Herramientas: R@FSOC

TAO

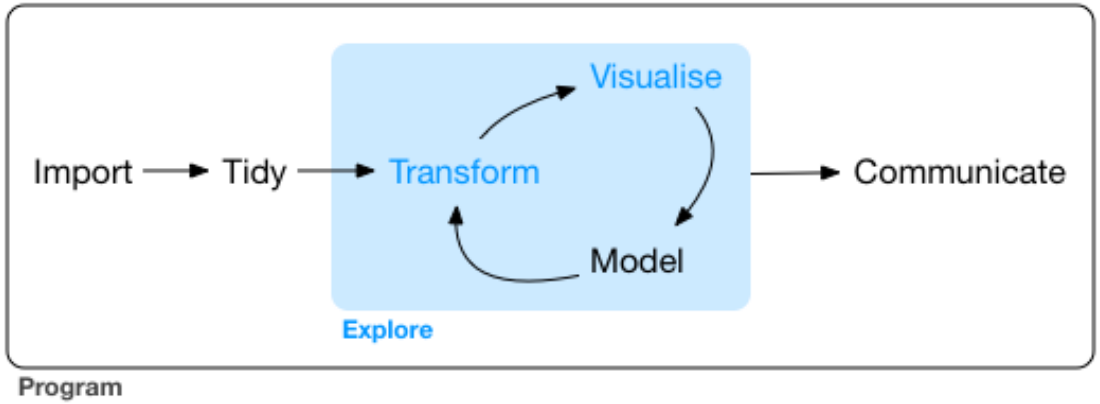
05/21/2021



Flujo de Trabajo

Flujo de Trabajo

Flujo de trabajo típico en C.D.



Manipulación Básica de Tablas de Datos

Librerías

```
library(tidyverse)
```

```
## -- 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()
```

DataFrames

Usamos datos y estructuras existentes en existentes en R: iris & Data Frame

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Del data.frame al tibble:

```
as_tibble(iris)
```

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

Tipo de Datos:

- *dbl* significa dobles, o números reales.
- *chr* significa vectores de caracteres o cadenas.
- *dttm* significa fechas y horas (una fecha + una hora).
- *lgl* significa lógico, vectores que solo contienen TRUE (verdadero) o FALSE (falso).
- *fctr* significa factores, que R usa para representar variables categóricas con valores posibles fijos.
- *date* significa fechas.

Creando un tible

```
tibble(
x = 1:5,
y = 1,
z = x^2 + y
)
```

```
## # A tibble: 5 x 3
##       x     y     z
##   <int> <dbl> <dbl>
## 1     1     1     2
## 2     2     1     5
## 3     3     1    10
## 4     4     1    17
## 5     5     1    26
```

Usando tribble para crear tibble

```
tribble(
~x, ~y, ~z,
"a", 2, 3.6,
"b", 1, 8.5
)
```

```
## # A tibble: 2 x 3
##   x         y     z
##   <chr> <dbl> <dbl>
## 1 a         2   3.6
## 2 b         1   8.5
```

Extraer Columnas

```
df <- tibble(
x = runif(5),
y = rnorm(5)
)
```

Extraer Columnas a Vector

Por nombre

```
df[["x"]]
```

```
## [1] 0.9951560 0.8960431 0.5834968 0.5076374 0.4413313
```

```
df$x
```

```
## [1] 0.9951560 0.8960431 0.5834968 0.5076374 0.4413313
```

Por posicion

```
df[[1]]
```

```
## [1] 0.9951560 0.8960431 0.5834968 0.5076374 0.4413313
```

Usando pipes

```
df %>% .$x
```

```
## [1] 0.9951560 0.8960431 0.5834968 0.5076374 0.4413313
```

```
df %>% .[["x"]]
```

```
## [1] 0.9951560 0.8960431 0.5834968 0.5076374 0.4413313
```

Extraer Columnas manteniendo el D.F.

Por nombre

```
df["x"]
```

```
## # A tibble: 5 x 1
##       x
##   <dbl>
## 1 0.995
## 2 0.896
## 3 0.583
## 4 0.508
## 5 0.441
```

Extraer Columnas manteniendo el D.F.

Por posicion

```
df[1]
```

```
## # A tibble: 5 x 1
##       x
##   <dbl>
## 1 0.995
## 2 0.896
## 3 0.583
## 4 0.508
## 5 0.441
```

Extraer Columnas manteniendo el D.F.

Usando pipes

```
df %>% .["x"]
```

```
## # A tibble: 5 x 1
##       x
##   <dbl>
## 1 0.995
## 2 0.896
## 3 0.583
## 4 0.508
## 5 0.441
```

Transformando Datos

DataSet de Juguete

```
flights = read_csv("vuelos.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   year = col_double(),
##   month = col_double(),
##   day = col_double(),
##   dep_time = col_double(),
##   sched_dep_time = col_double(),
##   dep_delay = col_double(),
##   arr_time = col_double(),
##   sched_arr_time = col_double(),
##   arr_delay = col_double(),
##   carrier = col_character(),
##   flight = col_double(),
##   tailnum = col_character(),
##   origin = col_character(),
##   dest = col_character(),
##   air_time = col_double(),
##   distance = col_double(),
##   hour = col_double(),
##   minute = col_double(),
##   time_hour = col_datetime(format = "")
## )
```

DataSet de Juguete

```
head(flights)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
## 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 <dbl>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Funciones para transformar y manipular datos

- filter()
- arrange()
- select()
- mutate()
- summarise()
- group_by()

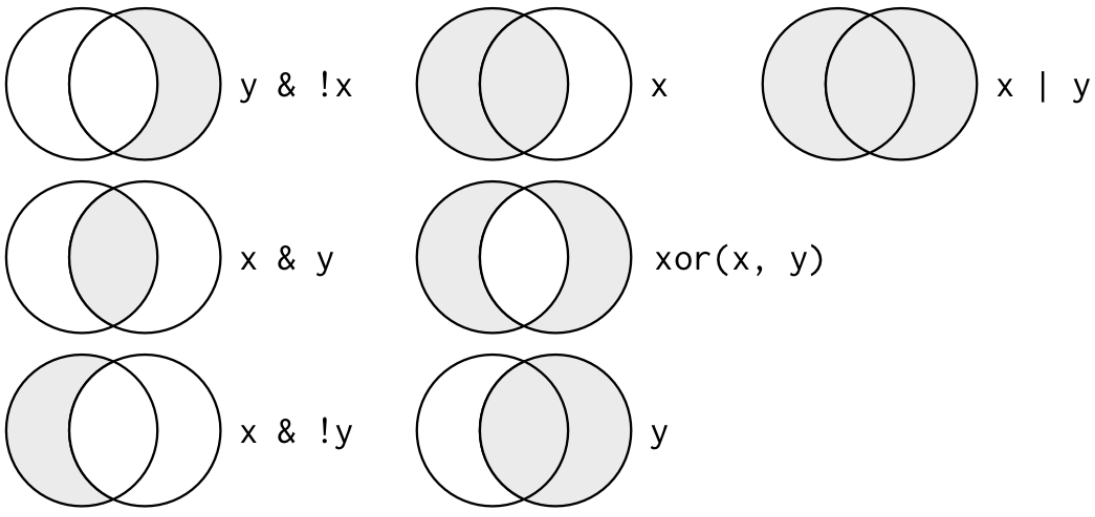
Filter

filter()

```
filter(flights, month == 1, day == 1)
```

```
## # A tibble: 842 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  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
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 832 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Operaciones lógicas



Más Ejemplos

```
filter(flights, (month == 11 | month == 12) & !(dep_delay <= 0) )
```

```
## # A tibble: 21,789 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  1  2013    11     1        5           2359         6      352            345
##  2  2013    11     1       35           2250       105      123           2356
##  3  2013    11     1      601            600         1      853            856
##  4  2013    11     1      602            600         2      843            815
##  5  2013    11     1      603            600         3      717            711
##  6  2013    11     1      623            600        23      806            758
##  7  2013    11     1      638            630         8      948            946
##  8  2013    11     1      639            635         4      830            833
##  9  2013    11     1      640            630        10      837            833
## 10  2013    11     1      651            640        11      812            807
## # ... with 21,779 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Más Ejemplos

```
filter(flights,  dep_delay < 60 )
```

```
## # A tibble: 301,462 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  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
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 301,452 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Cuidado con los NA's

Arrange

```
arrange(flights, year, month, day)
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  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
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Arrange

```
arrange(flights, desc(dep_delay))
```

```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  1  2013     1     9      641            900      1301     1242           1530
##  2  2013     6    15     1432           1935      1137     1607           2120
##  3  2013     1    10     1121           1635      1126     1239           1810
##  4  2013     9    20     1139           1845      1014     1457           2210
##  5  2013     7    22      845           1600      1005     1044           1815
##  6  2013     4    10     1100           1900       960     1342           2211
##  7  2013     3    17     2321            810       911      135           1020
##  8  2013     6    27      959           1900       899     1236           2226
##  9  2013     7    22     2257            759       898      121           1026
## 10  2013    12     5      756           1700       896     1058           2020
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Select

```
sub_fligths = select(flights, month, day, tailnum, origin, dest, dep_delay)
sub_fligths
```

```
## # A tibble: 336,776 x 6
##    month   day tailnum origin dest  dep_delay
##    <dbl> <dbl> <chr>   <chr>  <chr>     <dbl>
##  1     1     1 N14228  EWR    IAH           2
##  2     1     1 N24211  LGA    IAH           4
##  3     1     1 N619AA  JFK    MIA           2
##  4     1     1 N804JB  JFK    BQN          -1
##  5     1     1 N668DN  LGA    ATL          -6
##  6     1     1 N39463  EWR    ORD          -4
##  7     1     1 N516JB  EWR    FLL          -5
##  8     1     1 N829AS  LGA    IAD          -3
##  9     1     1 N593JB  JFK    MCO          -3
## 10     1     1 N3ALAA  LGA    ORD          -2
## # ... with 336,766 more rows
```

Select

```
# seleccionar cols excepto
select(flights, -(year:day))
```

```
## # A tibble: 336,776 x 16
##    dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
##       <dbl>          <dbl>     <dbl>    <dbl>          <dbl>     <dbl> <chr>
##  1      517            515         2      830            819        11 UA
##  2      533            529         4      850            830        20 UA
##  3      542            540         2      923            850        33 AA
##  4      544            545        -1     1004           1022       -18 B6
##  5      554            600        -6      812            837       -25 DL
##  6      554            558        -4      740            728        12 UA
##  7      555            600        -5      913            854        19 B6
##  8      557            600        -3      709            723       -14 EV
##  9      557            600        -3      838            846        -8 B6
## 10      558            600        -2      753            745         8 AA
## # ... with 336,766 more rows, and 9 more variables: flight <dbl>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Rename

```
rename(flights, año = year)
```

```
## # A tibble: 336,776 x 19
##      año month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <dbl> <dbl> <dbl>    <dbl>          <dbl>     <dbl>    <dbl>          <dbl>
##  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
##  7  2013     1     1      555            600        -5      913            854
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <dbl>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

Mutate

```
sub_fligths = mutate(sub_fligths, dep_puntual = dep_delay > 0, dep_puntual_grave = dep_delay > 30 )
sub_fligths
```

```
## # A tibble: 336,776 x 8
##    month   day tailnum origin dest  dep_delay dep_puntual dep_puntual_grave
##    <dbl> <dbl> <chr>   <chr>  <chr>     <dbl> <lgl>       <lgl>
##  1     1     1 N14228  EWR    IAH           2 TRUE        FALSE
##  2     1     1 N24211  LGA    IAH           4 TRUE        FALSE
##  3     1     1 N619AA  JFK    MIA           2 TRUE        FALSE
##  4     1     1 N804JB  JFK    BQN          -1 FALSE       FALSE
##  5     1     1 N668DN  LGA    ATL          -6 FALSE       FALSE
##  6     1     1 N39463  EWR    ORD          -4 FALSE       FALSE
##  7     1     1 N516JB  EWR    FLL          -5 FALSE       FALSE
##  8     1     1 N829AS  LGA    IAD          -3 FALSE       FALSE
##  9     1     1 N593JB  JFK    MCO          -3 FALSE       FALSE
## 10     1     1 N3ALAA  LGA    ORD          -2 FALSE       FALSE
## # ... with 336,766 more rows
```

Summarise

```
summarise(sub_fligths,
          delay = mean(dep_puntual, na.rm = TRUE),
          delay_grave = mean(dep_puntual_grave, na.rm = TRUE)
          )
```

```
## # A tibble: 1 x 2
##   delay delay_grave
##   <dbl>       <dbl>
## 1 0.391       0.147
```

Group By

```
agrupar_x_mes = group_by(sub_fligths,dest)
summarise(agrupar_x_mes,
          delay = mean(dep_puntual, na.rm = TRUE),
          delay_grave = mean(dep_puntual_grave, na.rm = TRUE),
          conteo = n()
          )
```

```
## # A tibble: 105 x 4
##    dest  delay delay_grave conteo
##    <chr> <dbl>       <dbl>  <int>
##  1 ABQ   0.453       0.169    254
##  2 ACK   0.309       0.109    265
##  3 ALB   0.501       0.272    439
##  4 ANC   0.75        0.125      8
##  5 ATL   0.360       0.134  17215
##  6 AUS   0.429       0.153   2439
##  7 AVL   0.327       0.133    275
##  8 BDL   0.451       0.228    443
##  9 BGR   0.425       0.256    375
## 10 BHM   0.507       0.324    297
## # ... with 95 more rows
```

Operador %>%

```
flights %>%
  select(month, day, tailnum, origin, dest, dep_delay) %>%
  filter(dep_delay < 60 ) %>%
  mutate(dep_puntual = dep_delay > 0, dep_puntual_grave = dep_delay > 30 ) %>%
  group_by(dest)  %>%
  summarise(
          delay = mean(dep_puntual, na.rm = TRUE),
          delay_grave = mean(dep_puntual_grave, na.rm = TRUE),
          conteo = n()
          )
```

```
## # A tibble: 104 x 4
##    dest  delay delay_grave conteo
##    <chr> <dbl>       <dbl>  <int>
##  1 ABQ   0.403      0.0944    233
##  2 ACK   0.277      0.0672    253
##  3 ALB   0.410      0.138     354
##  4 ANC   0.714      0           7
##  5 ATL   0.306      0.0615  15588
##  6 AUS   0.382      0.0833   2234
##  7 AVL   0.280      0.0732    246
##  8 BDL   0.372      0.117     360
##  9 BGR   0.330      0.133     309
## 10 BHM   0.394      0.167     221
## # ... with 94 more rows
```