Ciencia de Datos para Políticas Públicas

Módulo 2 - Clase 3: Transformación de datos/R Markdown

Pablo Aguirre Hörmann
29/06/2021

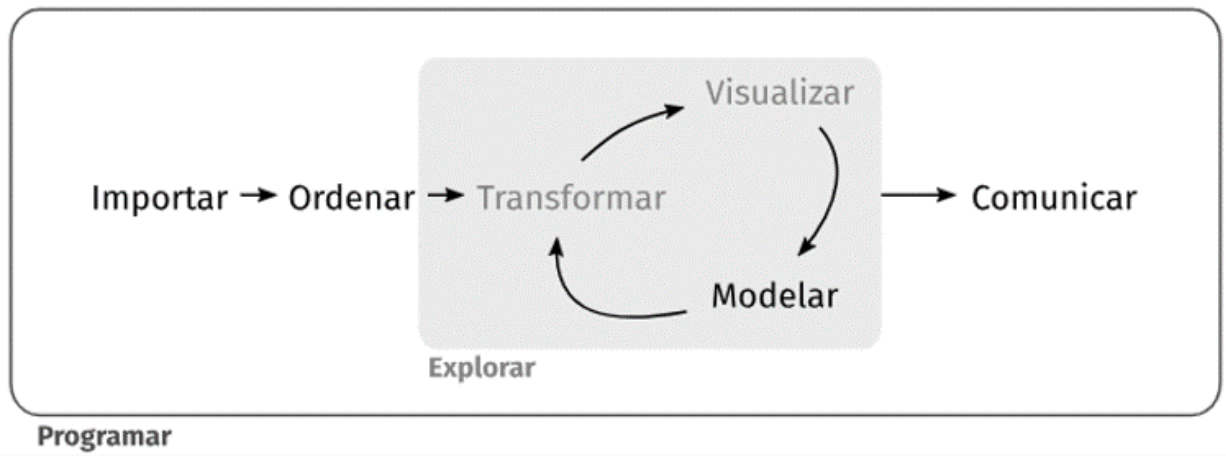
¿Qué veremos hoy?

- Visualización de datos
- Manejo de datos
- **Transformación de datos/** `R Markdown`
- Inferencia Estadística/Econometría

Pero antes...

- `EjercicioManejoDatosAlcaldes.R`

El contexto... nuevamente



Tidy data - Datos ordenados

Tidy data - Datos ordenados

- Cada columna es una variable
- Cada fila es una observación
- Cada celda corresponde a un valor

Tidyr - Cambio en funciones

- **Antes**: `spread`
- **Ahora**: `pivot_wider`

country	year	type	count
A	1999	cases	0.7K
A	1999	pop	19M
A	2000	cases	2K
A	2000	pop	20M
B	1999	cases	37K
B	1999	pop	172M
B	2000	cases	80K
B	2000	pop	174M
C	1999	cases	212K
C	1999	pop	1T
C	2000	cases	213K
C	2000	pop	1T

→

country	year	cases	pop
A	1999	0.7K	19M
A	2000	2K	20M
B	1999	37K	172M
B	2000	80K	174M
C	1999	212K	1T
C	2000	213K	1T

pivot_wider()

```r
table2
```

```
## # A tibble: 12 x 4
##    country      year type            count
##    <chr>       <int> <chr>           <int>
##  1 Afghanistan  1999 cases             745
##  2 Afghanistan  1999 population   19987071
##  3 Afghanistan  2000 cases            2666
##  4 Afghanistan  2000 population   20595360
##  5 Brazil       1999 cases           37737
##  6 Brazil       1999 population  172006362
##  7 Brazil       2000 cases           80488
##  8 Brazil       2000 population  174504898
##  9 China        1999 cases          212258
## 10 China        1999 population 1272915272
## 11 China        2000 cases          213766
## 12 China        2000 population 1280428583
```

```r
table2 %>%
  pivot_wider(names_from = type,
              values_from = count)
```

```
## # A tibble: 6 x 4
##   country      year  cases population
##   <chr>       <int>  <int>      <int>
## 1 Afghanistan  1999    745   19987071
## 2 Afghanistan  2000   2666   20595360
## 3 Brazil       1999  37737  172006362
## 4 Brazil       2000  80488  174504898
## 5 China        1999 212258 1272915272
## 6 China        2000 213766 1280428583
```

Tidyr - Cambio en funciones

- **Antes**: `gather`
- **Ahora**: `pivot_longer`

country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K

→

country	year	cases
A	1999	0.7K
B	1999	37K
C	1999	212K
A	2000	2K
B	2000	80K
C	2000	213K

pivot_longer()

```r
table4a
```

```
## # A tibble: 3 x 3
##   country     `1999` `2000`
## * <chr>        <int>  <int>
## 1 Afghanistan    745   2666
## 2 Brazil       37737  80488
## 3 China       212258 213766
```

```r
table4a %>%
  pivot_longer(2:3,
               names_to = "year",
               values_to = "value")
```

```
## # A tibble: 6 x 3
##   country     year   value
##   <chr>       <chr>  <int>
## 1 Afghanistan 1999     745
## 2 Afghanistan 2000    2666
## 3 Brazil      1999   37737
## 4 Brazil      2000   80488
## 5 China       1999  212258
## 6 China       2000  213766
```

Demo - Tuberculosis

script

- `Clase03.R`

Datos Tuberculosis

```r
glimpse(who)
```

```
## Rows: 7,240
## Columns: 60
## $ country      <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanis ...
## $ iso2         <chr> "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF",  ...
## $ iso3         <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG" ...
## $ year         <int> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988,  ...
## $ new_sp_m014  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m65   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f014  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f65   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m014  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m65   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
```

Datos Tuberculosis

```r
glimpse(who)
```

```
## Rows: 7,240
## Columns: 60
## $ country      <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanis ...
## $ iso2         <chr> "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF",  ...
## $ iso3         <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG" ...
## $ year         <int> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988,  ...
## $ new_sp_m014  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_m65   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f014  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sp_f65   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m014  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m1524 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m2534 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m3544 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m4554 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m5564 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
## $ new_sn_m65   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA ...
```

Donde comenzamos

```
(enfermedades <- who %>%
  select(-iso2, -iso3))
```

```
## # A tibble: 7,240 x 58
##    country  year new_sp_m014 new_sp_m1524 new_sp_m2534 new_sp_m3544 new_sp_m4554
##    <chr>   <int>       <int>        <int>        <int>        <int>        <int>
##  1 Afghan~  1980          NA           NA           NA           NA           NA
##  2 Afghan~  1981          NA           NA           NA           NA           NA
##  3 Afghan~  1982          NA           NA           NA           NA           NA
##  4 Afghan~  1983          NA           NA           NA           NA           NA
##  5 Afghan~  1984          NA           NA           NA           NA           NA
##  6 Afghan~  1985          NA           NA           NA           NA           NA
##  7 Afghan~  1986          NA           NA           NA           NA           NA
##  8 Afghan~  1987          NA           NA           NA           NA           NA
##  9 Afghan~  1988          NA           NA           NA           NA           NA
## 10 Afghan~  1989          NA           NA           NA           NA           NA
## # ... with 7,230 more rows, and 51 more variables: new_sp_m5564 <int>,
## #   new_sp_m65 <int>, new_sp_f014 <int>, new_sp_f1524 <int>,
## #   new_sp_f2534 <int>, new_sp_f3544 <int>, new_sp_f4554 <int>,
## #   new_sp_f5564 <int>, new_sp_f65 <int>, new_sn_m014 <int>,
## #   new_sn_m1524 <int>, new_sn_m2534 <int>, new_sn_m3544 <int>,
## #   new_sn_m4554 <int>, new_sn_m5564 <int>, new_sn_m65 <int>,
## #   new_sn_f014 <int>, new_sn_f1524 <int>, new_sn_f2534 <int>,
## #   new_sn_f3544 <int>, new_sn_f4554 <int>, new_sn_f5564 <int>,
## #   new_sn_f65 <int>, new_ep_m014 <int>, new_ep_m1524 <int>,
## #   new_ep_m2534 <int>, new_ep_m3544 <int>, new_ep_m4554 <int>,
## #   new_ep_m5564 <int>, new_ep_m65 <int>, new_ep_f014 <int>,
## #   new_ep_f1524 <int>, new_ep_f2534 <int>, new_ep_f3544 <int>,
## #   new_ep_f4554 <int>, new_ep_f5564 <int>, new_ep_f65 <int>,
```

Donde queremos llegar

```
tabla_final
```

```
## # A tibble: 2 x 8
## # Groups:   sexo [2]
##   sexo    `0-14` `15-24` `25-34` `35-44` `45-54` `55-64`  `65+`
##   <chr>    <int>   <int>   <int>   <int>   <int>   <int>  <int>
## 1 hombres  97051  406084  495242  478700  417188  325188 288063
## 2 mujeres  99738  320620  347398  260839  184791  136441 129468
```

- **¿Edad?**
- **¿Sexo?**

¿Qué información tenemos disponible?

```
names(enfermedades)
```

```
##  [1] "country"      "year"         "new_sp_m014"  "new_sp_m1524" "new_sp_m2534"
##  [6] "new_sp_m3544" "new_sp_m4554" "new_sp_m5564" "new_sp_m65"   "new_sp_f014"
## [11] "new_sp_f1524" "new_sp_f2534" "new_sp_f3544" "new_sp_f4554" "new_sp_f5564"
## [16] "new_sp_f65"   "new_sn_m014"  "new_sn_m1524" "new_sn_m2534" "new_sn_m3544"
## [21] "new_sn_m4554" "new_sn_m5564" "new_sn_m65"   "new_sn_f014"  "new_sn_f1524"
## [26] "new_sn_f2534" "new_sn_f3544" "new_sn_f4554" "new_sn_f5564" "new_sn_f65"
## [31] "new_ep_m014"  "new_ep_m1524" "new_ep_m2534" "new_ep_m3544" "new_ep_m4554"
## [36] "new_ep_m5564" "new_ep_m65"   "new_ep_f014"  "new_ep_f1524" "new_ep_f2534"
## [41] "new_ep_f3544" "new_ep_f4554" "new_ep_f5564" "new_ep_f65"   "newrel_m014"
## [46] "newrel_m1524" "newrel_m2534" "newrel_m3544" "newrel_m4554" "newrel_m5564"
## [51] "newrel_m65"   "newrel_f014"  "newrel_f1524" "newrel_f2534" "newrel_f3544"
## [56] "newrel_f4554" "newrel_f5564" "newrel_f65"
```

- **[*new o new_*]+[*método diagnóstico*]+[_]+[*sexo*]+[*Rango de Edad*]**

¿pivot_ ... longer o wider?

```
enfermedades
```

```
## # A tibble: 7,240 x 58
##    country  year new_sp_m014 new_sp_m1524 new_sp_m2534 new_sp_m3544
##    <chr>   <int>       <int>        <int>        <int>        <int>
##  1 Afghan~  1980          NA           NA           NA           NA
##  2 Afghan~  1981          NA           NA           NA           NA
##  3 Afghan~  1982          NA           NA           NA           NA
##  4 Afghan~  1983          NA           NA           NA           NA
##  5 Afghan~  1984          NA           NA           NA           NA
##  6 Afghan~  1985          NA           NA           NA           NA
##  7 Afghan~  1986          NA           NA           NA           NA
##  8 Afghan~  1987          NA           NA           NA           NA
##  9 Afghan~  1988          NA           NA           NA           NA
## 10 Afghan~  1989          NA           NA           NA           NA
## # ... with 7,230 more rows, and 51 more variables: new_sp_m5564 <int
## #   new_sp_m65 <int>, new_sp_f014 <int>, new_sp_f1524 <int>,
## #   new_sp_f2534 <int>, new_sp_f3544 <int>, new_sp_f4554 <int>,
## #   new_sp_f5564 <int>, new_sp_f65 <int>, new_sn_m014 <int>,
## #   new_sn_m1524 <int>, new_sn_m2534 <int>, new_sn_m3544 <int>,
## #   new_sn_m4554 <int>, new_sn_m5564 <int>, new_sn_m65 <int>,
## #   new_sn_f014 <int>, new_sn_f1524 <int>, new_sn_f2534 <int>,
## #   new_sn_f3544 <int>, new_sn_f4554 <int>, new_sn_f5564 <int>,
## #   new_sn_f65 <int>, new_ep_m014 <int>, new_ep_m1524 <int>,
## #   new_ep_m2534 <int>, new_ep_m3544 <int>, new_ep_m4554 <int>,
## #   new_ep_m5564 <int>, new_ep_m65 <int>, new_ep_f014 <int>,
## #   new_ep_f1524 <int>, new_ep_f2534 <int>, new_ep_f3544 <int>,
## #   new_ep_f4554 <int>, new_ep_f5564 <int>, new_ep_f65 <int>,
## #   newrel_m014 <int>, newrel_m1524 <int>, newrel_m2534 <int>,
## #   newrel_m3544 <int>, newrel_m4554 <int>, newrel_m5564 <int>,
## #   newrel_m65 <int>, newrel_f014 <int>, newrel_f1524 <int>,
## #   newrel_f2534 <int>, newrel_f3544 <int>, newrel_f4554 <int>,
## #   newrel_f5564 <int>, newrel_f65 <int>
```

Transformar la forma

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores")
```

```
## # A tibble: 405,440 x 4
##    country      year variables    valores
##    <chr>       <int> <chr>          <int>
##  1 Afghanistan  1980 new_sp_m014       NA
##  2 Afghanistan  1980 new_sp_m1524      NA
##  3 Afghanistan  1980 new_sp_m2534      NA
##  4 Afghanistan  1980 new_sp_m3544      NA
##  5 Afghanistan  1980 new_sp_m4554      NA
##  6 Afghanistan  1980 new_sp_m5564      NA
##  7 Afghanistan  1980 new_sp_m65        NA
##  8 Afghanistan  1980 new_sp_f014       NA
##  9 Afghanistan  1980 new_sp_f1524      NA
## 10 Afghanistan  1980 new_sp_f2534      NA
## # ... with 405,430 more rows
```

stringr

Donde estabamos

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores")
```

```
## # A tibble: 405,440 x 4
##    country      year variables    valores
##    <chr>       <int> <chr>          <int>
##  1 Afghanistan  1980 new_sp_m014       NA
##  2 Afghanistan  1980 new_sp_m1524      NA
##  3 Afghanistan  1980 new_sp_m2534      NA
##  4 Afghanistan  1980 new_sp_m3544      NA
##  5 Afghanistan  1980 new_sp_m4554      NA
##  6 Afghanistan  1980 new_sp_m5564      NA
##  7 Afghanistan  1980 new_sp_m65        NA
##  8 Afghanistan  1980 new_sp_f014       NA
##  9 Afghanistan  1980 new_sp_f1524      NA
## 10 Afghanistan  1980 new_sp_f2534      NA
## # ... with 405,430 more rows
```

Eliminaremos parte de "variables"

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new"))
```

```
## # A tibble: 405,440 x 4
##    country      year variables valores
##    <chr>       <int> <chr>       <int>
##  1 Afghanistan  1980 sp_m014        NA
##  2 Afghanistan  1980 sp_m1524       NA
##  3 Afghanistan  1980 sp_m2534       NA
##  4 Afghanistan  1980 sp_m3544       NA
##  5 Afghanistan  1980 sp_m4554       NA
##  6 Afghanistan  1980 sp_m5564       NA
##  7 Afghanistan  1980 sp_m65         NA
##  8 Afghanistan  1980 sp_f014        NA
##  9 Afghanistan  1980 sp_f1524       NA
## 10 Afghanistan  1980 sp_f2534       NA
## # ... with 405,430 more rows
```

Separamos la columna "variables"

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  separate(variables,
           into = c("enfermedad", "otro"),
           sep = "_")
```

```
## # A tibble: 405,440 x 5
##    country      year enfermedad otro  valores
##    <chr>       <int> <chr>      <chr>   <int>
##  1 Afghanistan  1980 sp         m014       NA
##  2 Afghanistan  1980 sp         m1524      NA
##  3 Afghanistan  1980 sp         m2534      NA
##  4 Afghanistan  1980 sp         m3544      NA
##  5 Afghanistan  1980 sp         m4554      NA
##  6 Afghanistan  1980 sp         m5564      NA
##  7 Afghanistan  1980 sp         m65        NA
##  8 Afghanistan  1980 sp         f014       NA
##  9 Afghanistan  1980 sp         f1524      NA
## 10 Afghanistan  1980 sp         f2534      NA
## # ... with 405,430 more rows
```

Separamos la columna "otro"

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  separate(variables,
           into = c("enfermedad", "otro"),
           sep = "_") %>%
  separate(otro,
           into = c("sexo", "edad"),
           sep =  1)
```

```
## # A tibble: 405,440 x 6
##    country      year enfermedad sexo  edad  valores
##    <chr>       <int> <chr>      <chr> <chr>   <int>
##  1 Afghanistan  1980 sp         m     014        NA
##  2 Afghanistan  1980 sp         m     1524       NA
##  3 Afghanistan  1980 sp         m     2534       NA
##  4 Afghanistan  1980 sp         m     3544       NA
##  5 Afghanistan  1980 sp         m     4554       NA
##  6 Afghanistan  1980 sp         m     5564       NA
##  7 Afghanistan  1980 sp         m     65         NA
##  8 Afghanistan  1980 sp         f     014        NA
##  9 Afghanistan  1980 sp         f     1524       NA
## 10 Afghanistan  1980 sp         f     2534       NA
## # ... with 405,430 more rows
```

Lo mismo usando paquete **stringr**

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  transmute(country, year,
    enfermedad = case_when(
      str_detect(variables, "rel") ~ str_sub(variables,
      TRUE ~ str_sub(variables, 1,2)),
    sexo = case_when(
      str_detect(variables, "m") ~ "m",
      TRUE ~ "f"),
    edad = str_extract(variables, "\\d+"),
    valores)
```

```
## # A tibble: 405,440 x 6
##    country      year enfermedad sexo  edad  valores
##    <chr>       <int> <chr>      <chr> <chr>   <int>
##  1 Afghanistan  1980 sp         m     014        NA
##  2 Afghanistan  1980 sp         m     1524       NA
##  3 Afghanistan  1980 sp         m     2534       NA
##  4 Afghanistan  1980 sp         m     3544       NA
##  5 Afghanistan  1980 sp         m     4554       NA
##  6 Afghanistan  1980 sp         m     5564       NA
##  7 Afghanistan  1980 sp         m     65         NA
##  8 Afghanistan  1980 sp         f     014        NA
##  9 Afghanistan  1980 sp         f     1524       NA
## 10 Afghanistan  1980 sp         f     2534       NA
## # ... with 405,430 more rows
```

¿transmute?

Casos (`case`) cada 10.000 personas

```
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  mutate(casos_pop = (cases/population)*10000)
```

```
## # A tibble: 6 x 5
##   country      year  cases population casos_pop
##   <chr>       <int>  <int>      <int>     <dbl>
## 1 Afghanistan  1999    745   19987071     0.373
## 2 Afghanistan  2000   2666   20595360     1.29
## 3 Brazil       1999  37737  172006362     2.19
## 4 Brazil       2000  80488  174504898     4.61
## 5 China        1999 212258 1272915272     1.67
## 6 China        2000 213766 1280428583     1.67
```

```
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  transmute(casos_pop = (cases/population)*10000)
```

```
## # A tibble: 6 x 1
##   casos_pop
##       <dbl>
## 1     0.373
## 2     1.29
## 3     2.19
## 4     4.61
## 5     1.67
## 6     1.67
```

¿case_when?

```r
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  mutate(indicador = ifelse(year == 1999,
                            1,
                            0))
```

```
## # A tibble: 6 x 5
##   country      year  cases population indicador
##   <chr>       <int>  <int>      <int>     <dbl>
## 1 Afghanistan  1999    745   19987071         1
## 2 Afghanistan  2000   2666   20595360         0
## 3 Brazil       1999  37737  172006362         1
## 4 Brazil       2000  80488  174504898         0
## 5 China        1999 212258 1272915272         1
## 6 China        2000 213766 1280428583         0
```

```r
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  mutate(indicador = case_when(
    year == 1999 ~ 1,
    year != 1999 ~ 0
  ))
```

```
## # A tibble: 6 x 5
##   country      year  cases population indicador
##   <chr>       <int>  <int>      <int>     <dbl>
## 1 Afghanistan  1999    745   19987071         1
## 2 Afghanistan  2000   2666   20595360         0
## 3 Brazil       1999  37737  172006362         1
## 4 Brazil       2000  80488  174504898         0
## 5 China        1999 212258 1272915272         1
## 6 China        2000 213766 1280428583         0
```

¿case_when?

```r
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  mutate(indicador = ifelse(year == 1999,
                            1,
                            0))
```

```
## # A tibble: 6 x 5
##   country      year  cases population indicador
##   <chr>       <int>  <int>      <int>     <dbl>
## 1 Afghanistan  1999    745   19987071         1
## 2 Afghanistan  2000   2666   20595360         0
## 3 Brazil       1999  37737  172006362         1
## 4 Brazil       2000  80488  174504898         0
## 5 China        1999 212258 1272915272         1
## 6 China        2000 213766 1280428583         0
```

```r
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  mutate(indicador = case_when(
    year == 1999 ~ 1,
    TRUE ~ 0
  ))
```

```
## # A tibble: 6 x 5
##   country      year  cases population indicador
##   <chr>       <int>  <int>      <int>     <dbl>
## 1 Afghanistan  1999    745   19987071         1
## 2 Afghanistan  2000   2666   20595360         0
## 3 Brazil       1999  37737  172006362         1
## 4 Brazil       2000  80488  174504898         0
## 5 China        1999 212258 1272915272         1
## 6 China        2000 213766 1280428583         0
```

¿case_when?

```r
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  mutate(indicador = ifelse(year == 1999,
                            1,
                            ifelse(country == "Brazil",
                                   2,
                                   0)))
```

```
## # A tibble: 6 x 5
##   country      year  cases population indicador
##   <chr>       <int>  <int>      <int>     <dbl>
## 1 Afghanistan  1999    745   19987071         1
## 2 Afghanistan  2000   2666   20595360         0
## 3 Brazil       1999  37737  172006362         1
## 4 Brazil       2000  80488  174504898         2
## 5 China        1999 212258 1272915272         1
## 6 China        2000 213766 1280428583         0
```

```r
table2 %>%
  pivot_wider(names_from = type,
              values_from = count) %>%
  mutate(indicador = case_when(
    year == 1999 ~ 1,
    country == "Brazil" ~ 2,
    TRUE ~ 0
  ))
```

```
## # A tibble: 6 x 5
##   country      year  cases population indicador
##   <chr>       <int>  <int>      <int>     <dbl>
## 1 Afghanistan  1999    745   19987071         1
## 2 Afghanistan  2000   2666   20595360         0
## 3 Brazil       1999  37737  172006362         1
## 4 Brazil       2000  80488  174504898         2
## 5 China        1999 212258 1272915272         1
## 6 China        2000 213766 1280428583         0
```

Expresiones regulares

```
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  transmute(country, year,
    enfermedad = case_when(
      str_detect(variables, "rel") ~ str_sub(variables,
      TRUE ~ str_sub(variables, 1,2)),
    sexo = case_when(
      str_detect(variables, "m") ~ "m",
      TRUE ~ "f"),
    edad = str_extract(variables, "\\d+"),
    valores)
```

- Las expresiones regulares son herramientas/instrucciones para describir patrones en texto
- Recursos:
  - https://stringr.tidyverse.org/articles/regular-expressions.html
  - http://griverorz.net/big-data/06-text-analysis/01-intro-regex.nb.html
  - https://rpubs.com/ydmarinb/429756
  - https://lost-stats.github.io/Data_Manipulation/Regular_Expressions.ht

Lo mismo usando paquete **stringr**

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  transmute(country, year,
    enfermedad = case_when(
      str_detect(variables, "rel") ~ str_sub(variables,
      TRUE ~ str_sub(variables, 1,2)),
    sexo = case_when(
      str_detect(variables, "m") ~ "m",
      TRUE ~ "f"),
    edad = str_extract(variables, "\\d+"),
    valores)
```

```
## # A tibble: 405,440 x 6
##    country      year enfermedad sexo  edad  valores
##    <chr>       <int> <chr>      <chr> <chr>   <int>
##  1 Afghanistan  1980 sp         m     014        NA
##  2 Afghanistan  1980 sp         m     1524       NA
##  3 Afghanistan  1980 sp         m     2534       NA
##  4 Afghanistan  1980 sp         m     3544       NA
##  5 Afghanistan  1980 sp         m     4554       NA
##  6 Afghanistan  1980 sp         m     5564       NA
##  7 Afghanistan  1980 sp         m     65         NA
##  8 Afghanistan  1980 sp         f     014        NA
##  9 Afghanistan  1980 sp         f     1524       NA
## 10 Afghanistan  1980 sp         f     2534       NA
## # ... with 405,430 more rows
```

Paso a paso

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  transmute(country, year,
    enfermedad = case_when(
      str_detect(variables, "rel") ~ str_sub(variables,
      TRUE ~ str_sub(variables, 1,2)),
    valores)
```

```
## # A tibble: 405,440 x 4
##    country      year enfermedad valores
##    <chr>       <int> <chr>        <int>
##  1 Afghanistan  1980 sp              NA
##  2 Afghanistan  1980 sp              NA
##  3 Afghanistan  1980 sp              NA
##  4 Afghanistan  1980 sp              NA
##  5 Afghanistan  1980 sp              NA
##  6 Afghanistan  1980 sp              NA
##  7 Afghanistan  1980 sp              NA
##  8 Afghanistan  1980 sp              NA
##  9 Afghanistan  1980 sp              NA
## 10 Afghanistan  1980 sp              NA
## # ... with 405,430 more rows
```

Paso a paso

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  transmute(country, year,
    enfermedad = case_when(
      str_detect(variables, "rel") ~ str_sub(variables,
      TRUE ~ str_sub(variables, 1,2)),
    sexo = case_when(
      str_detect(variables, "m") ~ "m",
      TRUE ~ "f"),
    valores)
```

```
## # A tibble: 405,440 x 5
##    country      year enfermedad sexo  valores
##    <chr>       <int> <chr>      <chr>   <int>
##  1 Afghanistan  1980 sp         m          NA
##  2 Afghanistan  1980 sp         m          NA
##  3 Afghanistan  1980 sp         m          NA
##  4 Afghanistan  1980 sp         m          NA
##  5 Afghanistan  1980 sp         m          NA
##  6 Afghanistan  1980 sp         m          NA
##  7 Afghanistan  1980 sp         m          NA
##  8 Afghanistan  1980 sp         f          NA
##  9 Afghanistan  1980 sp         f          NA
## 10 Afghanistan  1980 sp         f          NA
## # ... with 405,430 more rows
```

Paso a paso

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  transmute(country, year,
    enfermedad = case_when(
      str_detect(variables, "rel") ~ str_sub(variables,
      TRUE ~ str_sub(variables, 1,2)),
    sexo = case_when(
      str_detect(variables, "m") ~ "m",
      TRUE ~ "f"),
    edad = str_extract(variables, "\\d+"),
    valores)
```

```
## # A tibble: 405,440 x 6
##    country      year enfermedad sexo  edad  valores
##    <chr>       <int> <chr>      <chr> <chr>   <int>
##  1 Afghanistan  1980 sp         m     014        NA
##  2 Afghanistan  1980 sp         m     1524       NA
##  3 Afghanistan  1980 sp         m     2534       NA
##  4 Afghanistan  1980 sp         m     3544       NA
##  5 Afghanistan  1980 sp         m     4554       NA
##  6 Afghanistan  1980 sp         m     5564       NA
##  7 Afghanistan  1980 sp         m     65         NA
##  8 Afghanistan  1980 sp         f     014        NA
##  9 Afghanistan  1980 sp         f     1524       NA
## 10 Afghanistan  1980 sp         f     2534       NA
## # ... with 405,430 more rows
```

Mismo resultado

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  separate(variables,
           into = c("enfermedad", "otro"),
           sep = "_") %>%
  separate(otro,
           into = c("sexo", "edad"),
           sep =  1)
```

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  transmute(country, year,
    enfermedad = case_when(
      str_detect(variables, "rel") ~ str_sub(variables, 1, 3),
      TRUE ~ str_sub(variables, 1,2)),
    sexo = case_when(
      str_detect(variables, "m") ~ "m",
      TRUE ~ "f"),
    edad = str_extract(variables, "\\d+"),
    valores)
```

Cambios para más entendimiento

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  separate(variables,
           into = c("enfermedad", "otro"),
           sep = "_") %>%
  separate(otro,
           into = c("sexo", "edad"),
           sep =  1) %>%
  mutate(
    edad = case_when(
      edad == "014" ~ "0-14",
      edad == "1524" ~ "15-24",
      edad == "2534" ~ "25-34",
      edad == "3544" ~ "35-44",
      edad == "4554" ~ "45-54",
      edad == "5564" ~ "55-64",
      edad == "65" ~ "65+"),
    sexo = case_when(
      sexo == "m" ~ "hombres",
      sexo == "f" ~ "mujeres"))
```

```
## # A tibble: 405,440 x 6
##    country      year enfermedad sexo    edad  valores
##    <chr>       <int> <chr>      <chr>   <chr>   <int>
##  1 Afghanistan  1980 sp         hombres 0-14       NA
##  2 Afghanistan  1980 sp         hombres 15-24      NA
##  3 Afghanistan  1980 sp         hombres 25-34      NA
##  4 Afghanistan  1980 sp         hombres 35-44      NA
##  5 Afghanistan  1980 sp         hombres 45-54      NA
##  6 Afghanistan  1980 sp         hombres 55-64      NA
##  7 Afghanistan  1980 sp         hombres 65+        NA
##  8 Afghanistan  1980 sp         mujeres 0-14       NA
##  9 Afghanistan  1980 sp         mujeres 15-24      NA
## 10 Afghanistan  1980 sp         mujeres 25-34      NA
## # ... with 405,430 more rows
```

Total por sexo/edad para 2010

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  separate(variables,
           into = c("enfermedad", "otro"),
           sep = "_") %>%
  separate(otro,
           into = c("sexo", "edad"),
           sep =  1) %>%
  mutate(
    edad = case_when(
      edad == "014" ~ "0-14",
      edad == "1524" ~ "15-24",
      edad == "2534" ~ "25-34",
      edad == "3544" ~ "35-44",
      edad == "4554" ~ "45-54",
      edad == "5564" ~ "55-64",
      edad == "65" ~ "65+"),
    sexo = case_when(
      sexo == "m" ~ "hombres",
      sexo == "f" ~ "mujeres")) %>%
  filter(year == 2010) %>%
  group_by(sexo, edad) %>%
  summarise(total = sum(valores, na.rm = TRUE))
```

```
## # A tibble: 14 x 3
## # Groups:   sexo [2]
##    sexo    edad   total
##    <chr>   <chr>  <int>
##  1 hombres 0-14   97051
##  2 hombres 15-24 406084
##  3 hombres 25-34 495242
##  4 hombres 35-44 478700
##  5 hombres 45-54 417188
##  6 hombres 55-64 325188
##  7 hombres 65+   288063
##  8 mujeres 0-14   99738
##  9 mujeres 15-24 320620
## 10 mujeres 25-34 347398
## 11 mujeres 35-44 260839
## 12 mujeres 45-54 184791
## 13 mujeres 55-64 136441
## 14 mujeres 65+   129468
```

Tabla final

```r
enfermedades %>%
  pivot_longer(-c(country:year),
               names_to = "variables",
               values_to = "valores") %>%
  mutate(variables = str_remove(variables,
                                "new_"),
         variables = str_remove(variables,
                                "new")) %>%
  separate(variables,
           into = c("enfermedad", "otro"),
           sep = "_") %>%
  separate(otro,
           into = c("sexo", "edad"),
           sep =  1) %>%
  mutate(
    edad = case_when(
      edad == "014" ~ "0-14",
      edad == "1524" ~ "15-24",
      edad == "2534" ~ "25-34",
      edad == "3544" ~ "35-44",
      edad == "4554" ~ "45-54",
      edad == "5564" ~ "55-64",
      edad == "65" ~ "65+"),
    sexo = case_when(
      sexo == "m" ~ "hombres",
      sexo == "f" ~ "mujeres")) %>%
  filter(year == 2010) %>%
  group_by(sexo, edad) %>%
  summarise(total = sum(valores, na.rm = TRUE)) %>%
  pivot_wider(names_from = edad,
              values_from = total)
```

```
## # A tibble: 2 x 8
## # Groups:   sexo [2]
##   sexo    `0-14` `15-24` `25-34` `35-44` `45-54` `55-64`  `65+`
##   <chr>    <int>   <int>   <int>   <int>   <int>   <int>  <int>
## 1 hombres  97051  406084  495242  478700  417188  325188 288063
## 2 mujeres  99738  320620  347398  260839  184791  136441 129468
```

Ejercicio

Ejercicio

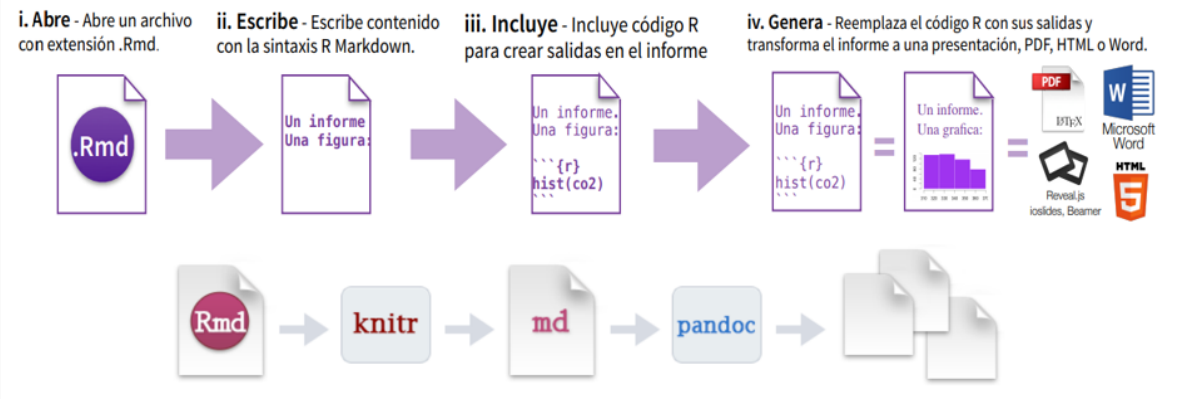
- `EjercicioManejoTransformacionDatos`

Reportería - R Markdown

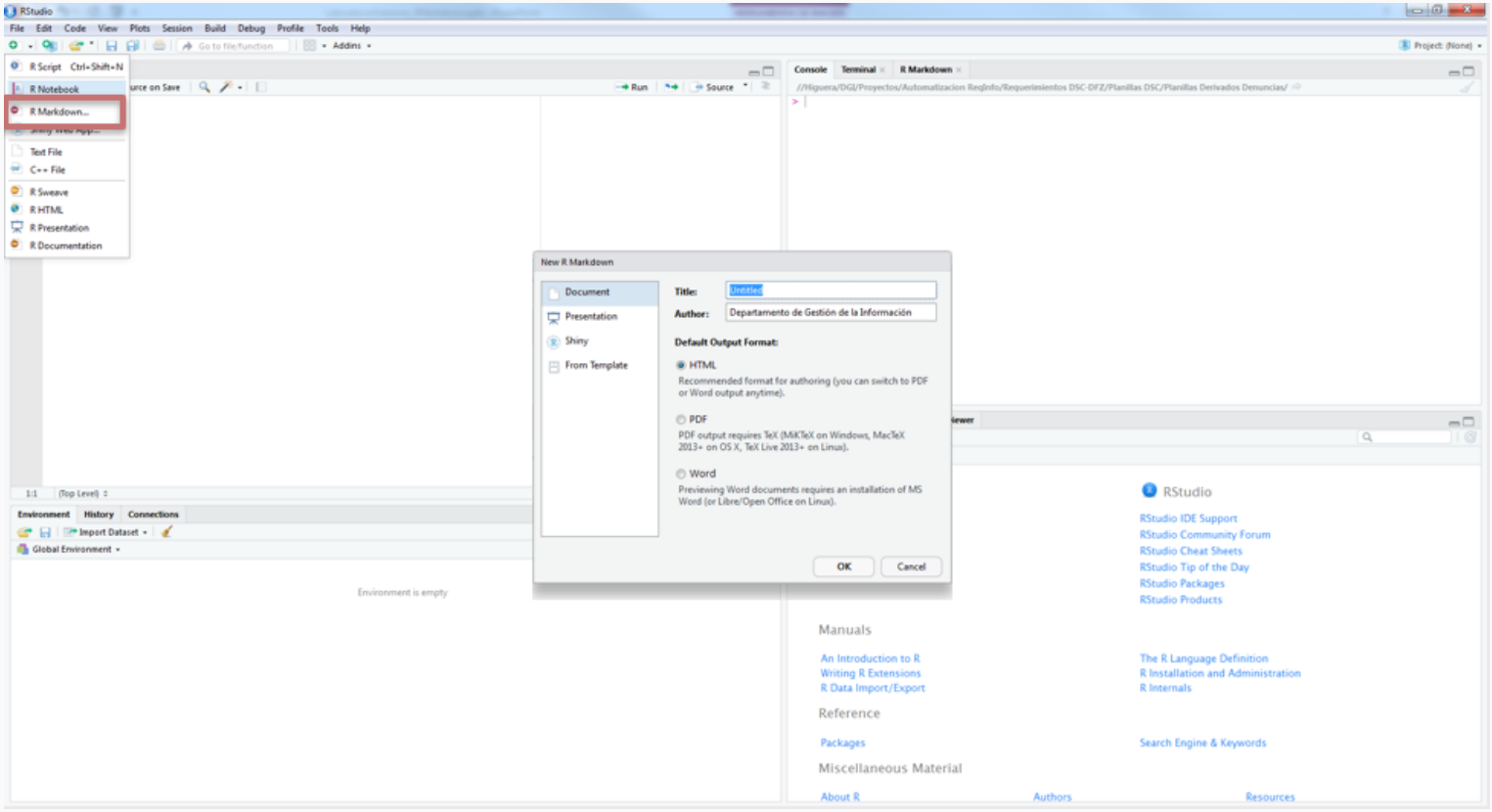
¿Qué es R Markdown?

Entorno para la creación de reportes/documentos reproducibles

- **HTML**
- MS Word
- PDF
- MS Power Point
- Y más...

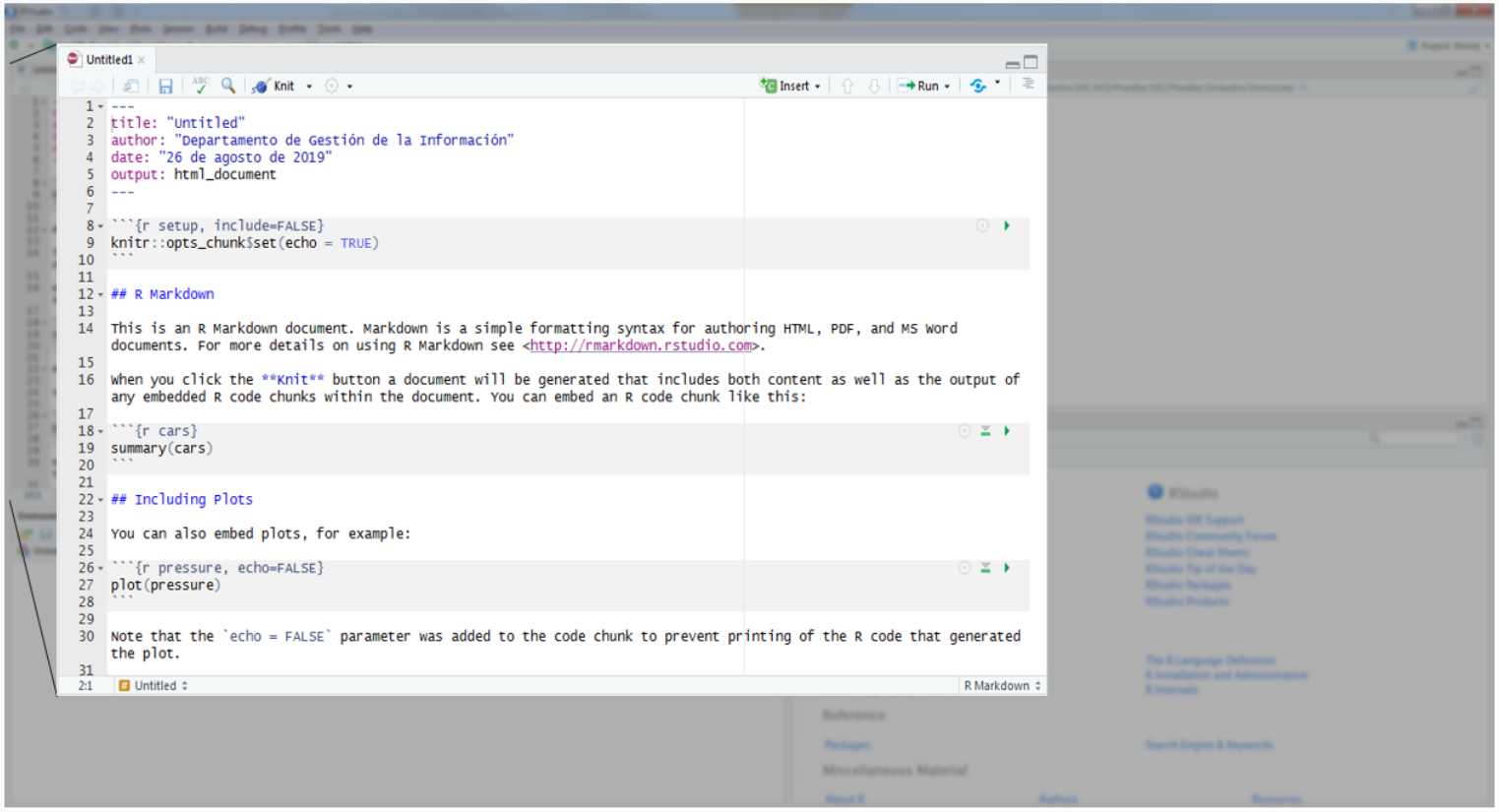


R Markdown



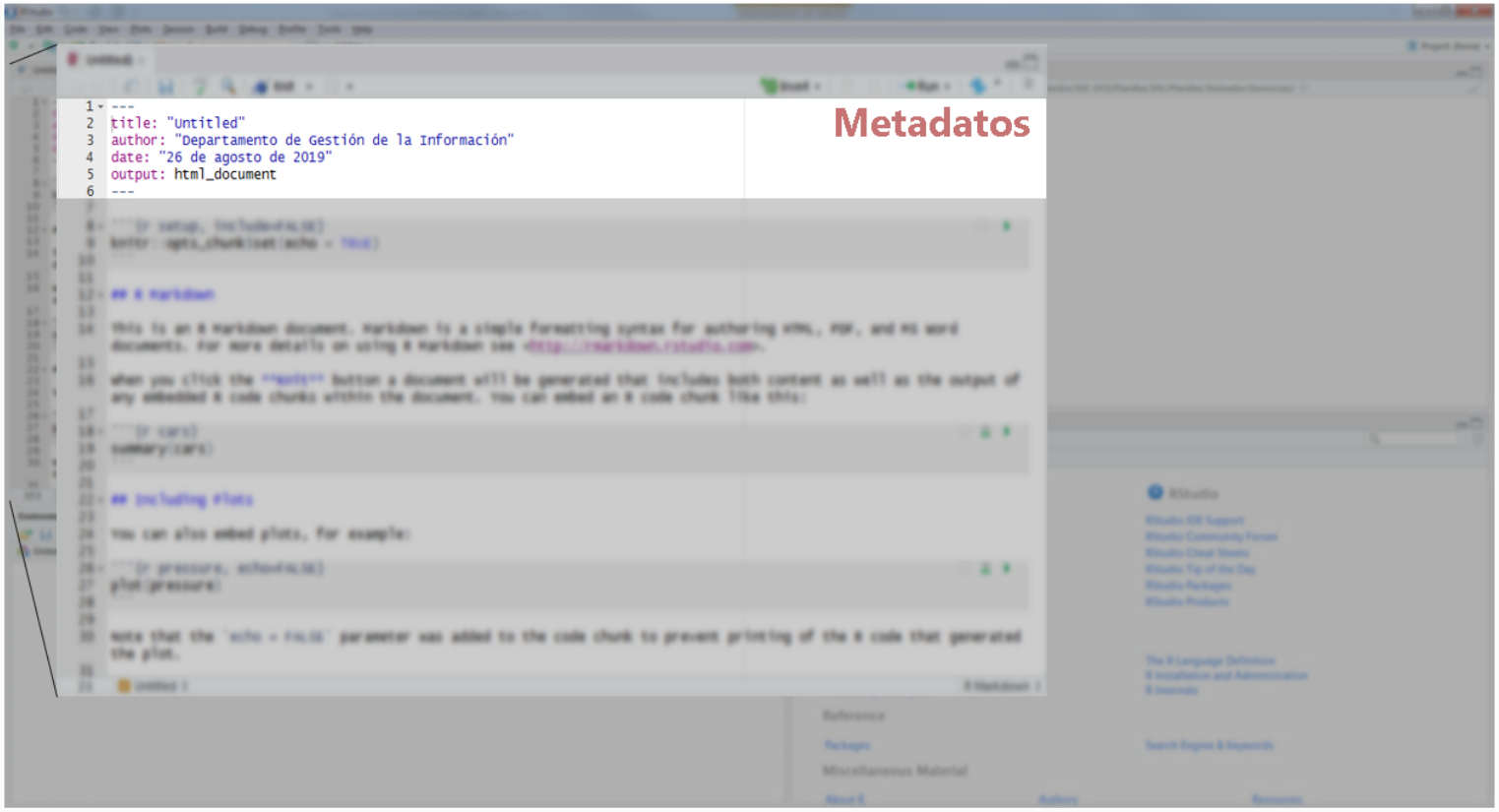
A diferencia de un script (`.R`), estos archivos serán `.Rmd`

R Markdown

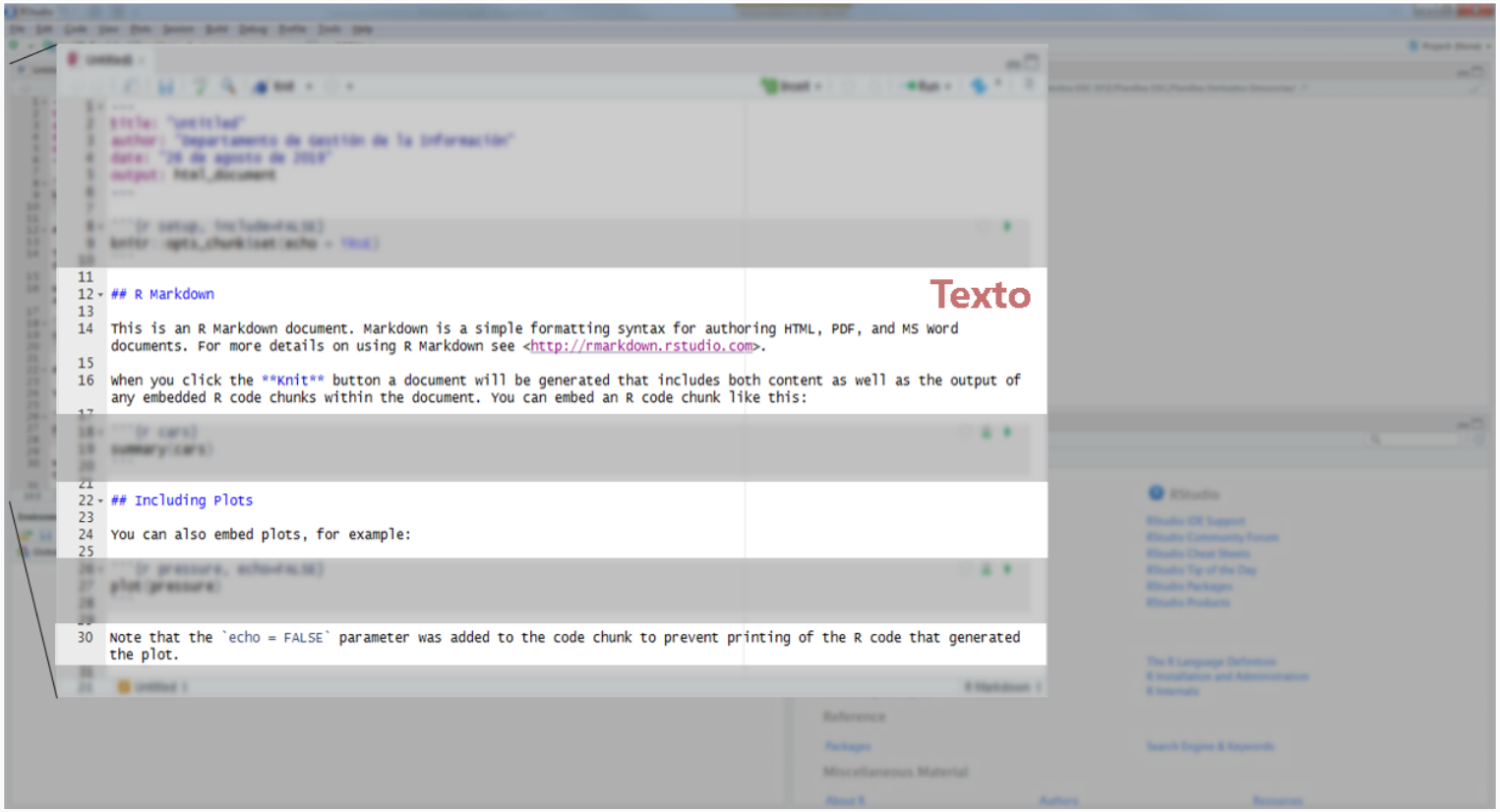

Untitled1
Knit
Insert
Run
---
title: "Untitled"
author: "Departamento de Gestión de la Información"
date: "26 de agosto de 2019"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
```{r cars}
summary(cars)
```
Including Plots
You can also embed plots, for example:
```{r pressure, echo=FALSE}
plot(pressure)
```
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
2:1
Untitled
R Markdown

R Markdown


Metadatos
---
title: "Untitled"
author: "Departamento de Gestión de la Información"
date: "26 de agosto de 2019"
output: html_document

R Markdown


Texto
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
Including Plots
You can also embed plots, for example:
Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Formato de texto

```
- *cursiva*
- **negrita**
- __*cursiva y negrita*__
- ~~tachado~~
- [link](https://gobierno.uai.cl/centro-investigacion/goblab-uai/)
- `objetos de código`
- entre otras..
```

- *cursiva*
- **negrita**
- ***cursiva y negrita***
- ~~tachado~~
- link
- `Objetos de código`
- entre otras..

Formato de texto

```
# Título 1
## Título 2
### Título 3
#### Título 4
##### Título 5
###### Título 6
```

Título 1

Título 2

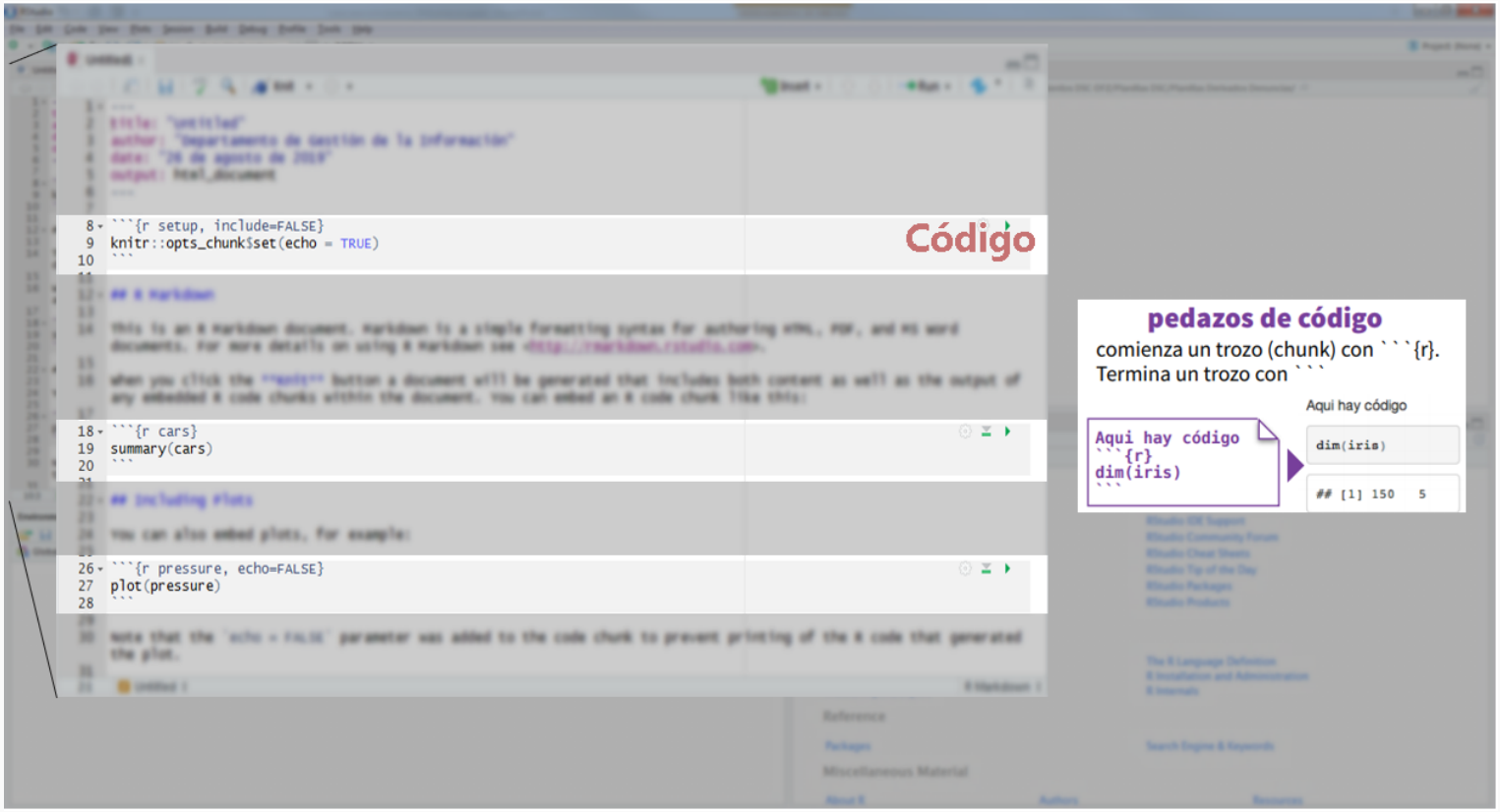
Título 3

Título 4

Título 5

Título 6

R Markdown


Código
pedazos de código
comienza un trozo (chunk) con ```{r}.
Termina un trozo con ```
Aqui hay código
Aqui hay código
```{r}
dim(iris)
```
dim(iris)
[1] 150 5

"Pedazos de código" (*chunks*)

- Se pueden agregar de distintas formas:
  - Ocupando las teclas **Ctrl**+**Alt**+**I** (Apple: **Cmd**+**Alt**+**I**).
  - Utilizando el botón *Insert new code* en la barra de tareas del archivo RMarkdown.
  - Escribiendo los delimitadores de los *chunks* manualmente: ```{r} y.
- Opciones de los *chunks* (se agregan en los encabezados de los *chunks* - ```{r})
  - `message = FALSE` previene que mensajes generados por el código aparezcan en el documento final.
  - `include = FALSE` previene que el código y los resultados del *chunk* aparezcan en el documento final. El código se ejecutará y los resultados estarán disponibles para ser usados en otros *chunks*.
  - `echo = FALSE` previene que el código aparezca en el resultado final pero el resultado si lo hará. Esto será muy útil cuando generemos tablas y/o gráficos
  - `warning = FALSE` previene que mensajes de error generados por el código aparezcan en el documento final.

Pueden revisar más opciones en la siguiente Hoja de Referencia.

Demostración

Demostración

- `DemoRMarkdown.Rmd`

Presentación 9 de Junio 2020

https://www.youtube.com/watch?v=QfeTzUF_8Nk

Lo que se viene

- Tarea 1 (11 de julio)
- Ayudantía (jueves)
- "Hora de consultas" (miércoles)

Próxima clase

- Inferencia estadísitca
- Intervalos de confianza
- Prueba de hipótesis
- Paquete `infer`