Ciencia de Datos para Políticas Públicas

Módulo 2 - Clase 2: Manejo de datos

Pablo Aguirre Hörmann
22/06/2021

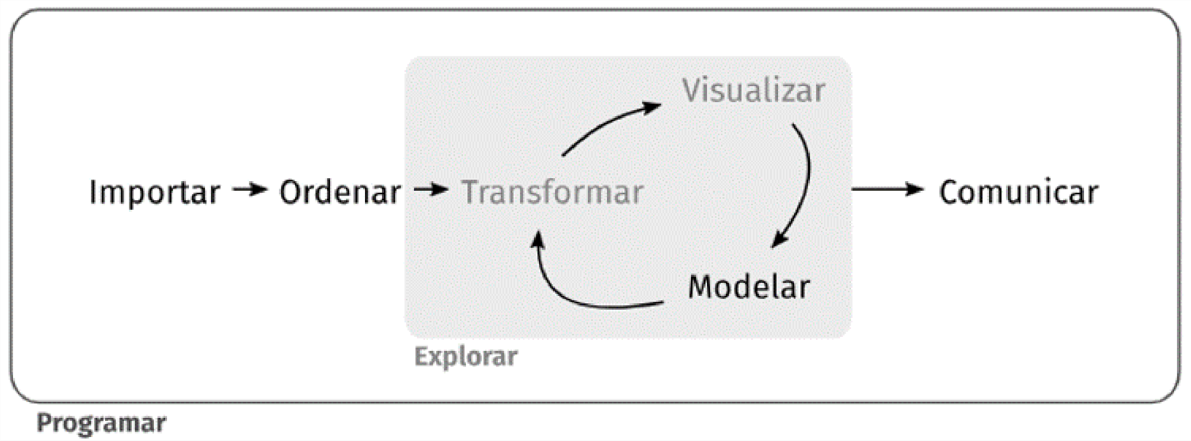
¿Qué veremos hoy?

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

Pero antes...

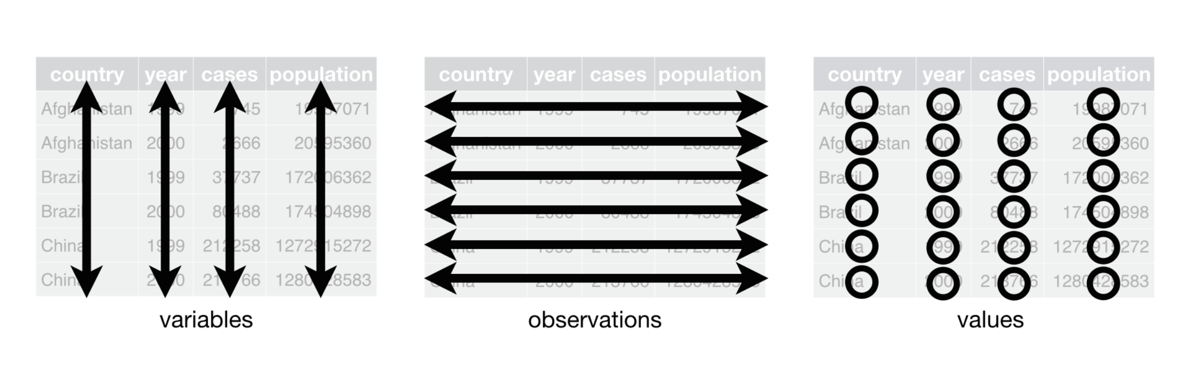
- `EjercicioGrafAlcaldes.R`

Recordemos el contexto



Datos *tidy* (ordenados)

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



Pipe

`%>%` nos permite definir nuestras acciones como una secuencia

- **Código "anidado"**

```r
estacionar(manejar(buscar(llaves), hacia = "trabajo"))
```

- **Código "por partes"**

```r
paso1 <- buscar(llaves)
paso2 <- manejar(paso1, hacia = "trabajo")
paso3 <- estacionar(paso2)
```

- **Código como secuencia**

```r
llaves %>%
  buscar() %>%
  manejar(hacia = "trabajo") %>%
  estacionar()
```

Pipe

```
log(sqrt(10))

paso1 <- sqrt(10)
paso2 <- log(paso1)

10 %>% sqrt() %>% log()
```

```
summary(iris)
iris %>% summary()
```

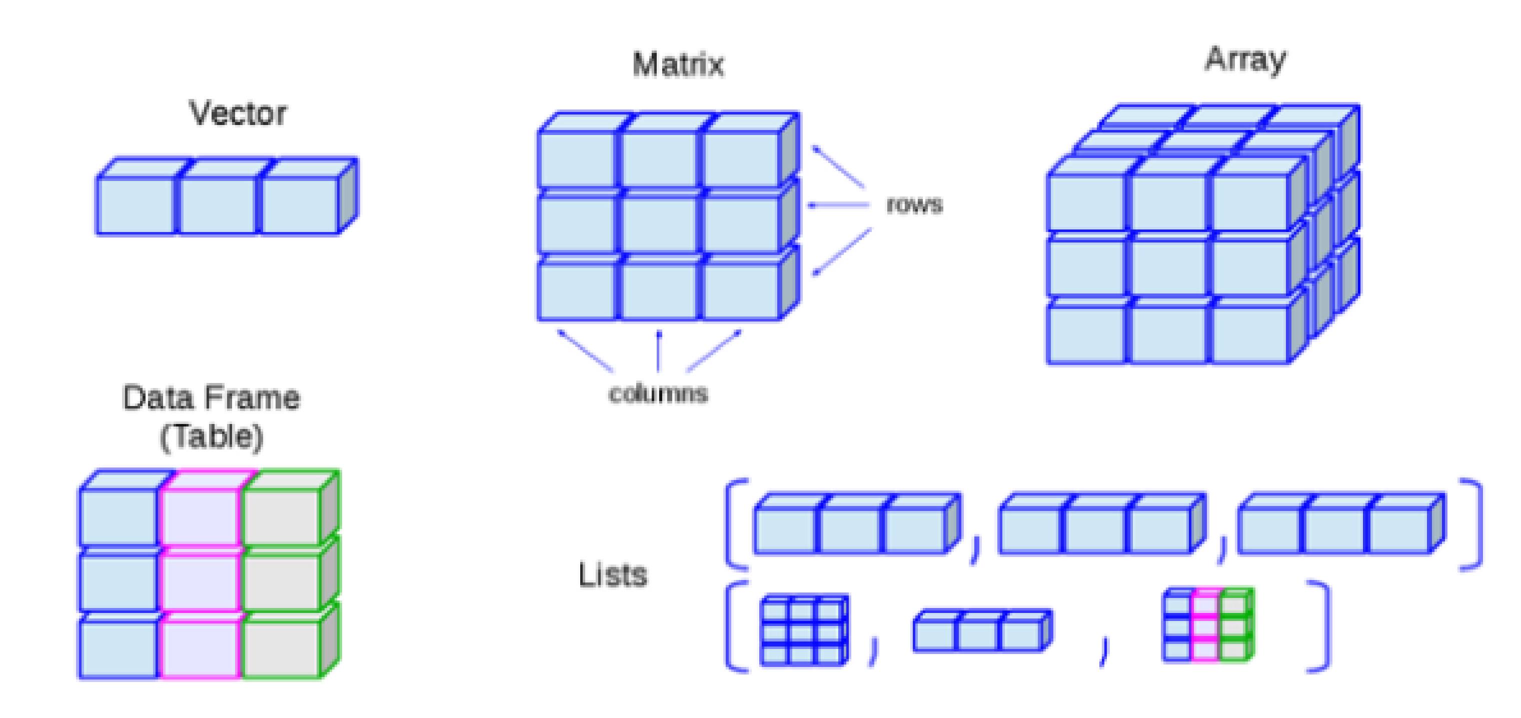
```
round(3.45, digits = 1)
3.45 %>% round(digits = 1)
```

Ojo

- No confundir `%>%` de `dplyr` con `+` de `ggplot2`
- `%>%` nos permite tomar un output y pasarlo/encadenarlo en la siguiente operación
- `+` nos permite crear capas en un gráfico

Manejo de datos

Foco en *data frames*



Manejo de datos - Script

- `ManejoDatos.R`

Datos ONU

```r
library(readr) # Cargar datos
library(dplyr) # Verbos de manipulación de datos
datosONU_tidy <- read_csv("../datos/DatosONU_tidy.csv")
names(datosONU_tidy)
```

```
##  [1] "country_name"
##  [2] "income_group"
##  [3] "region"
##  [4] "year"
##  [5] "co2_emissions_metric_tons_per_capita"
##  [6] "fertility_rate_total_births_per_woman"
##  [7] "forest_area_percent_of_land_area"
##  [8] "gdp_per_capita_constant_2005_us"
##  [9] "health_expenditure_per_capita_ppp_constant_2005_international"
## [10] "labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate"
## [11] "life_expectancy_at_birth_total_years"
## [12] "malnutrition_prevalence_weight_for_age_percent_of_children_under_5"
## [13] "population_total"
## [14] "urban_population_percent_of_total"
## [15] "fossil_fuel_energy_consumption_percent_of_total"
## [16] "poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population"
## [17] "public_spending_on_education_total_percent_of_government_expenditure"
```

Datos ONU

```
glimpse(datosONU_tidy)
```

```
## Rows: 7,704
## Columns: 12
## $ country_name                                    <chr> "Afghanistan", "Afg…
## $ income_group                                    <chr> "Low Income", "Low …
## $ region                                          <chr> "South Asia", "Sout…
## $ year                                            <dbl> 1972, 1973, 1974, 1…
## $ co2_emissions_metric_tons_per_capita            <dbl> 0.13163487, 0.13697…
## $ fertility_rate_total_births_per_woman           <dbl> 7.671, 7.671, 7.671…
## $ forest_area_percent_of_land_area                <dbl> NA, NA, NA, NA, NA,…
## $ gdp_per_capita_constant_2005_us                 <dbl> NA, NA, NA, NA, NA,…
## $ life_expectancy_at_birth_total_years            <dbl> 37.60888, 38.06934,…
## $ population_total                                <dbl> 11644377, 11966352,…
## $ urban_population_percent_of_total               <dbl> 11.9298, 12.3792, 1…
## $ fossil_fuel_energy_consumption_percent_of_total <dbl> NA, NA, NA, NA, NA,…
```

Funciones para manejo de datos

dplyr se basa en el concepto de funciones como verbos para manipular *data frames*

- `filter`: elige filas que cumplan criterio
- `slice`: elige filas según posición
- `select`: elige columnas según su nombre/posición
- `mutate`: crear nuevas columnas
- `rename`: cambio de nombre de columnas
- `arrange`: reordenar filas
- `distinct`: filtra valores únicos de filas
- `summarise`: reducir variables a valores
- … (muchas más)

Más información en la web del paquete

Reglas de dplyr para sus funciones

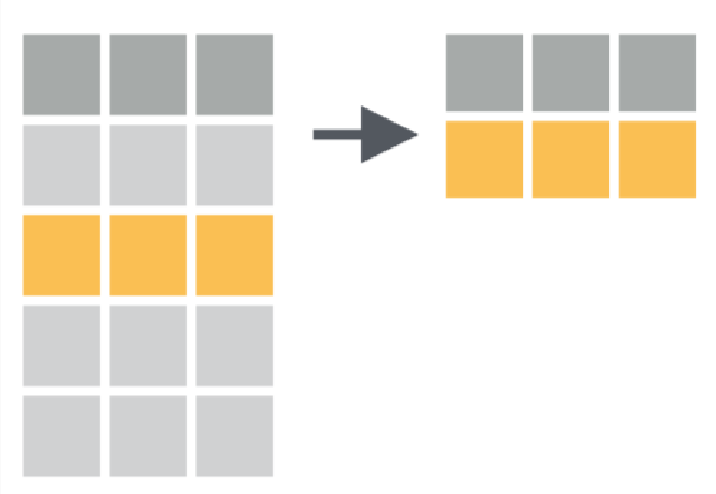
1. Primer argumento siempre es un *data frame*
2. Los siguientes argumentos describen que se hace con el *data frame*
3. El resultado siempre será un *data frame*

```
funcion(datos, instruccion1, instruccion2, ... )
```

filter()

Seleccionar filas

- `filter` permite seleccionar un subconjunto de filas de un *data frame*
  - Ej: filas donde la columna `X` es mayor a `n`.
- Se pueden poner muchas condiciones de forma simple



Seleccionar filas

Solo las observaciones correspondientes a Chile

```
datosONU_tidy %>%
  filter(country_name == "Chile")
```

```
## # A tibble: 36 x 17
##    country_name income_group region  year co2_emissions_m~ fertility_rate_~
##    <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
##  1 Chile        High Income  Latin~  1972             2.84             3.68
##  2 Chile        High Income  Latin~  1973             2.74             3.50
##  3 Chile        High Income  Latin~  1974             2.53             3.33
##  4 Chile        High Income  Latin~  1975             2.21             3.16
##  5 Chile        High Income  Latin~  1976             2.28             3.01
##  6 Chile        High Income  Latin~  1977             2.15             2.89
##  7 Chile        High Income  Latin~  1978             2.11             2.79
##  8 Chile        High Income  Latin~  1979             2.25             2.72
##  9 Chile        High Income  Latin~  1980             2.26             2.68
## 10 Chile        High Income  Latin~  1981             2.16             2.66
## # ... with 26 more rows, and 11 more variables:
## #   forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Seleccionar filas

Solo las observaciones correspondientes a Chile y para años posteriores al 2000

```
datosONU_tidy %>%
  filter(country_name == "Chile", year > 2000)
```

```
## # A tibble: 7 x 17
##   country_name income_group region  year co2_emissions_m~ fertility_rate_~
##   <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
## 1 Chile        High Income  Latin~  2001             3.37             2.05
## 2 Chile        High Income  Latin~  2002             3.50             2.01
## 3 Chile        High Income  Latin~  2003             3.44             1.98
## 4 Chile        High Income  Latin~  2004             3.71             1.96
## 5 Chile        High Income  Latin~  2005             3.78             1.94
## 6 Chile        High Income  Latin~  2006             3.90             1.92
## 7 Chile        High Income  Latin~  2007             4.27             1.90
## # … with 11 more variables: forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Seleccionar filas

Solo las observaciones correspondientes al 2000 o al 2007

```
datosONU_tidy %>%
  filter(year == 2000 | year == 2007)
```

```
## # A tibble: 428 x 17
##    country_name income_group region  year co2_emissions_m~ fertility_rate_~
##    <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
##  1 Afghanistan  Low Income   South~  2000           0.0379             7.73
##  2 Afghanistan  Low Income   South~  2007           0.0756             6.46
##  3 Albania      Upper Middl~ Europ~  2000           0.978              2.38
##  4 Albania      Upper Middl~ Europ~  2007           1.38               1.80
##  5 Algeria      Upper Middl~ Middl~  2000           2.77               2.51
##  6 Algeria      Upper Middl~ Middl~  2007           3.20               2.66
##  7 American Sa~ Upper Middl~ East ~  2000          NA                 NA
##  8 American Sa~ Upper Middl~ East ~  2007          NA                 NA
##  9 Andorra      High Income  Europ~  2000           8.02              NA
## 10 Andorra      High Income  Europ~  2007           6.63               1.18
## # ... with 418 more rows, and 11 more variables:
## #   forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Operadores lógicos

Operador	Definición
<	menor
<=	menor o igual
>	mayor
>=	mayor o igual
==	estrictamente igual
!=	distinto
x|y	x O y
x&y	x Y y

Operador	Definición
is.na(x)	test: valor NA (nulo)
!is.na(x)	x perteneciente a y
x %in% y	x perteneciente a y
!(x %in% y)	todo lo perteneciente a y que no es x
!x	no x

Seleccionar filas

Solo las observaciones correspondientes a los años 1995, 2000, y 2005

```
datosONU_tidy %>%
  filter(year == 1995 | year == 2000 | year == 2005)
```

```
## # A tibble: 642 x 17
##    country_name income_group region  year co2_emissions_m~ fertility_rate_~
##    <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
##  1 Afghanistan  Low Income   South~  1995           0.0721             7.83
##  2 Afghanistan  Low Income   South~  2000           0.0379             7.73
##  3 Afghanistan  Low Income   South~  2005           0.0409             6.93
##  4 Albania      Upper Middl~ Europ~  1995           0.655              2.72
##  5 Albania      Upper Middl~ Europ~  2000           0.978              2.38
##  6 Albania      Upper Middl~ Europ~  2005           1.42               1.92
##  7 Algeria      Upper Middl~ Middl~  1995           3.23               3.45
##  8 Algeria      Upper Middl~ Middl~  2000           2.77               2.51
##  9 Algeria      Upper Middl~ Middl~  2005           3.15               2.51
## 10 American Sa~ Upper Middl~ East ~  1995          NA                 NA
## # ... with 632 more rows, and 11 more variables:
## #   forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Seleccionar filas

Solo las observaciones correspondientes a los años 1995, 2000, y 2005

```r
datosONU_tidy %>%
  filter(year %in% c(1995, 2000, 2005))
```

```
## # A tibble: 642 x 17
##    country_name income_group region  year co2_emissions_m~ fertility_rate_~
##    <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
##  1 Afghanistan  Low Income   South~  1995           0.0721             7.83
##  2 Afghanistan  Low Income   South~  2000           0.0379             7.73
##  3 Afghanistan  Low Income   South~  2005           0.0409             6.93
##  4 Albania      Upper Middl~ Europ~  1995           0.655              2.72
##  5 Albania      Upper Middl~ Europ~  2000           0.978              2.38
##  6 Albania      Upper Middl~ Europ~  2005           1.42               1.92
##  7 Algeria      Upper Middl~ Middl~  1995           3.23               3.45
##  8 Algeria      Upper Middl~ Middl~  2000           2.77               2.51
##  9 Algeria      Upper Middl~ Middl~  2005           3.15               2.51
## 10 American Sa~ Upper Middl~ East ~  1995          NA                 NA
## # ... with 632 more rows, and 11 more variables:
## #   forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Seleccionar filas

Solo las observaciones **NO** correspondientes a los años 1995, 2000, y 2005

```
datosONU_tidy %>%
  filter(!year %in% c(1995, 2000, 2005))
```

```
## # A tibble: 7,062 x 17
##    country_name income_group region  year co2_emissions_m~ fertility_rate_~
##    <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
##  1 Afghanistan  Low Income   South~  1972            0.132             7.67
##  2 Afghanistan  Low Income   South~  1973            0.137             7.67
##  3 Afghanistan  Low Income   South~  1974            0.156             7.67
##  4 Afghanistan  Low Income   South~  1975            0.169             7.67
##  5 Afghanistan  Low Income   South~  1976            0.155             7.67
##  6 Afghanistan  Low Income   South~  1977            0.183             7.67
##  7 Afghanistan  Low Income   South~  1978            0.164             7.67
##  8 Afghanistan  Low Income   South~  1979            0.169             7.67
##  9 Afghanistan  Low Income   South~  1980            0.134             7.67
## 10 Afghanistan  Low Income   South~  1981            0.153             7.67
## # … with 7,052 more rows, and 11 more variables:
## #   forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Seleccionar filas por posición

La quinta fila

```
datosONU_tidy %>%
  slice(5)
```

```
## # A tibble: 1 x 17
##   country_name income_group region  year co2_emissions_m~ fertility_rate_~
##   <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
## 1 Afghanistan  Low Income   South~  1976            0.155             7.67
## # … with 11 more variables: forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Seleccionar filas por posición

Las primeras cinco filas

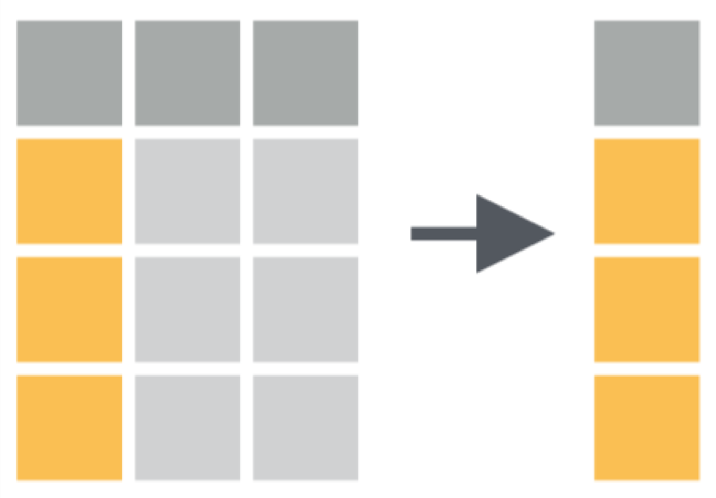
```
datosONU_tidy %>%
  slice(1:5)
```

```
## # A tibble: 5 x 17
##   country_name income_group region  year co2_emissions_m~ fertility_rate_~
##   <chr>        <chr>        <chr>  <dbl>            <dbl>            <dbl>
## 1 Afghanistan  Low Income   South~  1972            0.132             7.67
## 2 Afghanistan  Low Income   South~  1973            0.137             7.67
## 3 Afghanistan  Low Income   South~  1974            0.156             7.67
## 4 Afghanistan  Low Income   South~  1975            0.169             7.67
## 5 Afghanistan  Low Income   South~  1976            0.155             7.67
## #  … with 11 more variables: forest_area_percent_of_land_area <dbl>,
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

select()

Seleccionar columnas/variables

- `select` permite seleccionar un subconjunto de columnas de un *data frame*
  - U ordenarlas de una forma en particular
- Se pueden seleccionar por nombre o por posición



Seleccionar columnas/variables

Seleccionar 5 variables/columnas

```
datosONU_tidy %>%
  select(country_name, income_group, region, year, population_total)
```

```
## # A tibble: 7,704 x 5
##    country_name income_group region      year population_total
##    <chr>        <chr>        <chr>      <dbl>            <dbl>
##  1 Afghanistan  Low Income   South Asia  1972         11644377
##  2 Afghanistan  Low Income   South Asia  1973         11966352
##  3 Afghanistan  Low Income   South Asia  1974         12273589
##  4 Afghanistan  Low Income   South Asia  1975         12551790
##  5 Afghanistan  Low Income   South Asia  1976         12806810
##  6 Afghanistan  Low Income   South Asia  1977         13034460
##  7 Afghanistan  Low Income   South Asia  1978         13199597
##  8 Afghanistan  Low Income   South Asia  1979         13257128
##  9 Afghanistan  Low Income   South Asia  1980         13180431
## 10 Afghanistan  Low Income   South Asia  1981         12963788
## # … with 7,694 more rows
```

Seleccionar columnas/variables

Dejar todas las columnas menos dos

```
datosONU_tidy %>%
  select(-region, -income_group)
```

```
## # A tibble: 7,704 x 15
##    country_name  year co2_emissions_m~ fertility_rate_~ forest_area_per~
##    <chr>        <dbl>            <dbl>            <dbl>            <dbl>
##  1 Afghanistan   1972            0.132             7.67               NA
##  2 Afghanistan   1973            0.137             7.67               NA
##  3 Afghanistan   1974            0.156             7.67               NA
##  4 Afghanistan   1975            0.169             7.67               NA
##  5 Afghanistan   1976            0.155             7.67               NA
##  6 Afghanistan   1977            0.183             7.67               NA
##  7 Afghanistan   1978            0.164             7.67               NA
##  8 Afghanistan   1979            0.169             7.67               NA
##  9 Afghanistan   1980            0.134             7.67               NA
## 10 Afghanistan   1981            0.153             7.67               NA
## # ... with 7,694 more rows, and 10 more variables:
## #   gdp_per_capita_constant_2005_us <dbl>,
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

Funciones "de ayuda"

Dejar todas las columnas que contengan *per capita*

```
datosONU_tidy %>%
  select(contains("per_capita"))
```

```
## # A tibble: 7,704 x 3
##    co2_emissions_metric_t~ gdp_per_capita_const~ health_expenditure_per_capita_~
##                      <dbl>                 <dbl>                           <dbl>
##   1                  0.132                    NA                              NA
##   2                  0.137                    NA                              NA
##   3                  0.156                    NA                              NA
##   4                  0.169                    NA                              NA
##   5                  0.155                    NA                              NA
##   6                  0.183                    NA                              NA
##   7                  0.164                    NA                              NA
##   8                  0.169                    NA                              NA
##   9                  0.134                    NA                              NA
## 10                   0.153                    NA                              NA
## # … with 7,694 more rows
```

Funciones "de ayuda"

Dejar todas las columnas que contengan *per capita* o *poverty*

```
datosONU_tidy %>%
  select(contains("per_capita") | contains("poverty"))
```

```
## # A tibble: 7,704 x 4
##    co2_emissions_met~ gdp_per_capita_co~ health_expenditure~ poverty_headcount_~
##                 <dbl>              <dbl>               <dbl>               <dbl>
##  1              0.132                 NA                  NA                  NA
##  2              0.137                 NA                  NA                  NA
##  3              0.156                 NA                  NA                  NA
##  4              0.169                 NA                  NA                  NA
##  5              0.155                 NA                  NA                  NA
##  6              0.183                 NA                  NA                  NA
##  7              0.164                 NA                  NA                  NA
##  8              0.169                 NA                  NA                  NA
##  9              0.134                 NA                  NA                  NA
## 10              0.153                 NA                  NA                  NA
## # ... with 7,694 more rows
```

Funciones "de ayuda"

Dejar todas las columnas que contengan *per capita* o *poverty*

```
datosONU_tidy %>%
  select(contains(c("per_capita", "poverty")))
```

```
## # A tibble: 7,704 x 4
##    co2_emissions_met~ gdp_per_capita_co~ health_expenditure~ poverty_headcount_~
##                 <dbl>              <dbl>               <dbl>               <dbl>
##  1              0.132                 NA                  NA                  NA
##  2              0.137                 NA                  NA                  NA
##  3              0.156                 NA                  NA                  NA
##  4              0.169                 NA                  NA                  NA
##  5              0.155                 NA                  NA                  NA
##  6              0.183                 NA                  NA                  NA
##  7              0.164                 NA                  NA                  NA
##  8              0.169                 NA                  NA                  NA
##  9              0.134                 NA                  NA                  NA
## 10              0.153                 NA                  NA                  NA
## # … with 7,694 more rows
```

Funciones "de ayuda"

Dejar todas las columnas que contengan *per capita* y *co2*

```
datosONU_tidy %>%
  select(contains("per_capita") & contains("co2"))
```

```
## # A tibble: 7,704 x 1
##    co2_emissions_metric_tons_per_capita
##                                   <dbl>
##  1                                0.132
##  2                                0.137
##  3                                0.156
##  4                                0.169
##  5                                0.155
##  6                                0.183
##  7                                0.164
##  8                                0.169
##  9                                0.134
## 10                                0.153
## # … with 7,694 more rows
```

Funciones "de ayuda"

Dejar todas las columnas que comiencen con *p*

```
datosONU_tidy %>%
  select(starts_with("p"))
```

```
## # A tibble: 7,704 x 3
##    population_total poverty_headcount_ratio_at_2~ public_spending_on_education_~
##               <dbl>                         <dbl>                          <dbl>
##  1         11644377                            NA                             NA
##  2         11966352                            NA                             NA
##  3         12273589                            NA                             NA
##  4         12551790                            NA                             NA
##  5         12806810                            NA                             NA
##  6         13034460                            NA                             NA
##  7         13199597                            NA                             NA
##  8         13257128                            NA                             NA
##  9         13180431                            NA                             NA
## 10         12963788                            NA                             NA
## # … with 7,694 more rows
```

Funciones "de ayuda"

Dejar todas las columnas numéricas

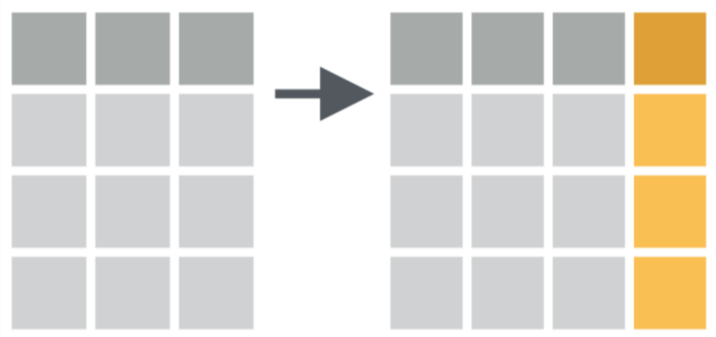
```
datosONU_tidy %>%
  select(where(is.numeric))
```

```
## # A tibble: 7,704 x 14
##     year co2_emissions_m~ fertility_rate_~ forest_area_per~ gdp_per_capita_~
##    <dbl>            <dbl>            <dbl>            <dbl>            <dbl>
##   1  1972            0.132             7.67               NA               NA
##   2  1973            0.137             7.67               NA               NA
##   3  1974            0.156             7.67               NA               NA
##   4  1975            0.169             7.67               NA               NA
##   5  1976            0.155             7.67               NA               NA
##   6  1977            0.183             7.67               NA               NA
##   7  1978            0.164             7.67               NA               NA
##   8  1979            0.169             7.67               NA               NA
##   9  1980            0.134             7.67               NA               NA
## 10  1981            0.153             7.67               NA               NA
## # … with 7,694 more rows, and 9 more variables:
## #   health_expenditure_per_capita_ppp_constant_2005_international <dbl>,
## #   labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate <dbl>,
## #   life_expectancy_at_birth_total_years <dbl>,
## #   malnutrition_prevalence_weight_for_age_percent_of_children_under_5 <dbl>,
## #   population_total <dbl>, urban_population_percent_of_total <dbl>,
## #   fossil_fuel_energy_consumption_percent_of_total <dbl>,
## #   poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population <dbl>,
## #   public_spending_on_education_total_percent_of_government_expenditure <dbl>
```

mutate()

Crear columnas/variables

- `mutate` permite generar nuevas columnas/variables en un *data frame*
  - Ej: nueva columna `Z` igual a la división entre las columnas `X` e `Y`
- Nuevas columnas pueden o no depender de columnas ya existentes
- Se pueden generar más de una columna en un comando



Crear columnas/variables

Nueva columna calculando el logaritmo de una existente

```
datosONU_tidy %>%
  select(country_name, year, co2_emissions_metric_tons_per_capita)
```

```
## # A tibble: 7,704 x 3
##    country_name  year co2_emissions_metric_tons_per_capita
##    <chr>        <dbl>                                <dbl>
##  1 Afghanistan   1972                                0.132
##  2 Afghanistan   1973                                0.137
##  3 Afghanistan   1974                                0.156
##  4 Afghanistan   1975                                0.169
##  5 Afghanistan   1976                                0.155
##  6 Afghanistan   1977                                0.183
##  7 Afghanistan   1978                                0.164
##  8 Afghanistan   1979                                0.169
##  9 Afghanistan   1980                                0.134
## 10 Afghanistan   1981                                0.153
## # … with 7,694 more rows
```

Crear columnas/variables

Nueva columna calculando el logaritmo de una existente

```
datosONU_tidy %>%
  select(country_name, year, co2_emissions_metric_tons_per_capita) %>%
  mutate(log_co2_emissions = log(co2_emissions_metric_tons_per_capita))
```

```
## # A tibble: 7,704 x 4
##    country_name  year co2_emissions_metric_tons_per_capita log_co2_emissions
##    <chr>        <dbl>                                <dbl>             <dbl>
##  1 Afghanistan   1972                                0.132             -2.03
##  2 Afghanistan   1973                                0.137             -1.99
##  3 Afghanistan   1974                                0.156             -1.86
##  4 Afghanistan   1975                                0.169             -1.78
##  5 Afghanistan   1976                                0.155             -1.86
##  6 Afghanistan   1977                                0.183             -1.70
##  7 Afghanistan   1978                                0.164             -1.81
##  8 Afghanistan   1979                                0.169             -1.78
##  9 Afghanistan   1980                                0.134             -2.01
## 10 Afghanistan   1981                                0.153             -1.88
## # … with 7,694 more rows
```

Cambiar nombres de columnas

Nombres muy largos

```
names(datosONU_tidy)
```

```
##  [1] "country_name"
##  [2] "income_group"
##  [3] "region"
##  [4] "year"
##  [5] "co2_emissions_metric_tons_per_capita"
##  [6] "fertility_rate_total_births_per_woman"
##  [7] "forest_area_percent_of_land_area"
##  [8] "gdp_per_capita_constant_2005_us"
##  [9] "health_expenditure_per_capita_ppp_constant_2005_international"
## [10] "labor_force_participation_rate_female_percent_of_female_population_ages_15_modeled_ilo_estimate"
## [11] "life_expectancy_at_birth_total_years"
## [12] "malnutrition_prevalence_weight_for_age_percent_of_children_under_5"
## [13] "population_total"
## [14] "urban_population_percent_of_total"
## [15] "fossil_fuel_energy_consumption_percent_of_total"
## [16] "poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population"
## [17] "public_spending_on_education_total_percent_of_government_expenditure"
```

Cambiar nombres de columnas

`rename` - estructura a seguir

```
rename(datos, NuevoNombre = AntiguoNombre)
```

Cambiar nombres de columnas

```
datosONU_tidy %>%
  rename(
    "co2_emissions"             = "co2_emissions_metric_tons_per_capita",
    "fertility_rate"            = "fertility_rate_total_births_per_woman",
    "forest_area"               = "forest_area_percent_of_land_area",
    "gdp_per_capita"            = "gdp_per_capita_constant_2005_us",
    "health_expenditure"        = "health_expenditure_per_capita_ppp_constant_2005_international",
    "labor_force_participation" = "labor_force_participation_rate_female_percent_of_female_population_ages_15_
    "life_expectancy"           = "life_expectancy_at_birth_total_years",
    "malnutrition_prevalence"   = "malnutrition_prevalence_weight_for_age_percent_of_children_under_5",
    "urban_population"          = "urban_population_percent_of_total",
    "fossil_fuel_consumption"   = "fossil_fuel_energy_consumption_percent_of_total",
    "poverty"                   = "poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population",
    "public_spending_education" = "public_spending_on_education_total_percent_of_government_expenditure"
  )
```

No olvidar "guardar" los resultados

Sobreescribir *data frame*

```
datosONU_tidy ← datosONU_tidy %>%
  rename(
    "co2_emissions"             = "co2_emissions_metric_tons_per_capita",
    "fertility_rate"            = "fertility_rate_total_births_per_woman",
    "forest_area"               = "forest_area_percent_of_land_area",
    "gdp_per_capita"            = "gdp_per_capita_constant_2005_us",
    "health_expenditure"        = "health_expenditure_per_capita_ppp_constant_2005_international",
    "labor_force_participation" = "labor_force_participation_rate_female_percent_of_female_population_ages_15_
    "life_expectancy"           = "life_expectancy_at_birth_total_years",
    "malnutrition_prevalence"   = "malnutrition_prevalence_weight_for_age_percent_of_children_under_5",
    "urban_population"          = "urban_population_percent_of_total",
    "fossil_fuel_consumption"   = "fossil_fuel_energy_consumption_percent_of_total",
    "poverty"                   = "poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population",
    "public_spending_education" = "public_spending_on_education_total_percent_of_government_expenditure"
  )
```

No olvidar "guardar" los resultados

Crear nuevo *data frame*

```
datosONU_tidy_nuevo <- datosONU_tidy %>%
  rename(
    "co2_emissions"              = "co2_emissions_metric_tons_per_capita",
    "fertility_rate"             = "fertility_rate_total_births_per_woman",
    "forest_area"                = "forest_area_percent_of_land_area",
    "gdp_per_capita"             = "gdp_per_capita_constant_2005_us",
    "health_expenditure"         = "health_expenditure_per_capita_ppp_constant_2005_international",
    "labor_force_participation" = "labor_force_participation_rate_female_percent_of_female_population_ages_15_
    "life_expectancy"            = "life_expectancy_at_birth_total_years",
    "malnutrition_prevalence"    = "malnutrition_prevalence_weight_for_age_percent_of_children_under_5",
    "urban_population"           = "urban_population_percent_of_total",
    "fossil_fuel_consumption"    = "fossil_fuel_energy_consumption_percent_of_total",
    "poverty"                    = "poverty_headcount_ratio_at_2_a_day_ppp_percent_of_population",
    "public_spending_education" = "public_spending_on_education_total_percent_of_government_expenditure"
  )
```

Revisar resultado del cambio de nombre

```
names(datosONU_tidy)
```

```
##  [1] "country_name"              "income_group"
##  [3] "region"                    "year"
##  [5] "co2_emissions"             "fertility_rate"
##  [7] "forest_area"               "gdp_per_capita"
##  [9] "health_expenditure"        "labor_force_participation"
## [11] "life_expectancy"           "malnutrition_prevalence"
## [13] "population_total"          "urban_population"
## [15] "fossil_fuel_consumption"   "poverty"
## [17] "public_spending_education"
```

summarise() / group_by()

Reducir variables a valores

Número de observaciones

```
datosONU_tidy %>%
  summarise(n_observaciones = n())
```

```
## # A tibble: 1 x 1
##   n_observaciones
##             <int>
## 1            7704
```

Número de países

```
datosONU_tidy %>%
  summarise(n_paises = n_distinct(country_name))
```

```
## # A tibble: 1 x 1
##   n_paises
##      <int>
## 1      214
```

Reducir variables a valores

Promedio de la columna *fertility_rate*

```
datosONU_tidy %>%
  summarise(promedio_fertility_rate = mean(fertility_rate, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   promedio_fertility_rate
##                     <dbl>
## 1                    3.95
```

Máximo valor de *gdp_per_capita*

```
datosONU_tidy %>%
  summarise(max_gdp_per_capita = max(gdp_per_capita, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   max_gdp_per_capita
##                <dbl>
## 1            147141.
```

Reducir variables a valores

Se puede calcular más de un valor a la vez

```
datosONU_tidy %>%
  summarise(n_observaciones         = n(),
            n_paises                = n_distinct(country_name),
            promedio_fertility_rate = mean(fertility_rate, na.rm = TRUE),
            max_gdp_per_capita      = max(gdp_per_capita, na.rm = TRUE))
```

```
## # A tibble: 1 x 4
##   n_observaciones n_paises promedio_fertility_rate max_gdp_per_capita
##             <int>    <int>                   <dbl>              <dbl>
## 1            7704      214                    3.95            147141.
```

Agrupar observaciones

Por si sola no pasa nada

```
datosONU_tidy %>%
  group_by(region)
```

```
## # A tibble: 7,704 x 17
## # Groups:   region [7]
##    country_name income_group region  year co2_emissions fertility_rate
##    <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
##  1 Afghanistan  Low Income   South~  1972         0.132           7.67
##  2 Afghanistan  Low Income   South~  1973         0.137           7.67
##  3 Afghanistan  Low Income   South~  1974         0.156           7.67
##  4 Afghanistan  Low Income   South~  1975         0.169           7.67
##  5 Afghanistan  Low Income   South~  1976         0.155           7.67
##  6 Afghanistan  Low Income   South~  1977         0.183           7.67
##  7 Afghanistan  Low Income   South~  1978         0.164           7.67
##  8 Afghanistan  Low Income   South~  1979         0.169           7.67
##  9 Afghanistan  Low Income   South~  1980         0.134           7.67
## 10 Afghanistan  Low Income   South~  1981         0.153           7.67
## # … with 7,694 more rows, and 11 more variables: forest_area <dbl>,
## #   gdp_per_capita <dbl>, health_expenditure <dbl>,
## #   labor_force_participation <dbl>, life_expectancy <dbl>,
## #   malnutrition_prevalence <dbl>, population_total <dbl>,
## #   urban_population <dbl>, fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Agrupar observaciones

Pero con `summarise` aparecen las ventajas

```
datosONU_tidy %>%
  group_by(region) %>%
  summarise(n_observaciones = n())
```

```
## # A tibble: 7 x 2
##   region                          n_observaciones
##   <chr>                                     <int>
## 1 East Asia and Pacific                      1296
## 2 Europe and Central Afica                   2052
## 3 Latin America and the Caribbean            1476
## 4 Middle East and North Africa                756
## 5 North America                               108
## 6 South Asia                                  288
## 7 Sub-saharan Africa                         1728
```

Agrupar observaciones

Pero con `summarise` aparecen las ventajas

```r
datosONU_tidy %>%
  group_by(region) %>%
  summarise(n_observaciones = n(),
            n_paises = n_distinct(country_name),
            promedio_fertility_rate = mean(fertility_rate, na.rm = TRUE))
```

```
## # A tibble: 7 x 4
##   region                         n_observaciones n_paises promedio_fertility_ra~
##   <chr>                                    <int>    <int>                  <dbl>
## 1 East Asia and Pacific                     1296       36                   3.60
## 2 Europe and Central Afica                  2052       57                   2.12
## 3 Latin America and the Caribbe~            1476       41                   3.42
## 4 Middle East and North Africa               756       21                   4.72
## 5 North America                              108        3                   1.80
## 6 South Asia                                 288        8                   5.04
## 7 Sub-saharan Africa                        1728       48                   6.09
```

Agrupar observaciones

Se puede agrupar por más de una variable/columna

```
datosONU_tidy %>%
  group_by(region, income_group) %>%
  summarise(n_paises = n_distinct(country_name))
```

```
## # A tibble: 24 x 3
## # Groups:   region [7]
##    region                          income_group        n_paises
##    <chr>                           <chr>                  <int>
##  1 East Asia and Pacific           High Income               13
##  2 East Asia and Pacific           Low Income                 1
##  3 East Asia and Pacific           Lower Middle Income       13
##  4 East Asia and Pacific           Upper Middle Income        9
##  5 Europe and Central Afica        High Income               36
##  6 Europe and Central Afica        Low Income                 1
##  7 Europe and Central Afica        Lower Middle Income        4
##  8 Europe and Central Afica        Upper Middle Income       16
##  9 Latin America and the Caribbean High Income               16
## 10 Latin America and the Caribbean Low Income                 1
## # … with 14 more rows
```

Otras funciones

Ordenar filas según columnas

```
datosONU_tidy
```

```
## # A tibble: 7,704 x 17
##    country_name income_group region  year co2_emissions fertility_rate
##    <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
##  1 Afghanistan  Low Income   South~  1972         0.132           7.67
##  2 Afghanistan  Low Income   South~  1973         0.137           7.67
##  3 Afghanistan  Low Income   South~  1974         0.156           7.67
##  4 Afghanistan  Low Income   South~  1975         0.169           7.67
##  5 Afghanistan  Low Income   South~  1976         0.155           7.67
##  6 Afghanistan  Low Income   South~  1977         0.183           7.67
##  7 Afghanistan  Low Income   South~  1978         0.164           7.67
##  8 Afghanistan  Low Income   South~  1979         0.169           7.67
##  9 Afghanistan  Low Income   South~  1980         0.134           7.67
## 10 Afghanistan  Low Income   South~  1981         0.153           7.67
## # … with 7,694 more rows, and 11 more variables: forest_area <dbl>,
## #   gdp_per_capita <dbl>, health_expenditure <dbl>,
## #   labor_force_participation <dbl>, life_expectancy <dbl>,
## #   malnutrition_prevalence <dbl>, population_total <dbl>,
## #   urban_population <dbl>, fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Ordenar filas según columnas

Ordenar según la columna *year*

```
datosONU_tidy %>%
  arrange(year)
```

```
## # A tibble: 7,704 x 17
##    country_name income_group region  year co2_emissions fertility_rate
##    <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
##  1 Afghanistan  Low Income   South~  1972         0.132           7.67
##  2 Albania      Upper Middl~ Europ~  1972         2.52            4.81
##  3 Algeria      Upper Middl~ Middl~  1972         1.83            7.59
##  4 American Sa~ Upper Middl~ East ~  1972        NA              NA
##  5 Andorra      High Income  Europ~  1972        NA              NA
##  6 Angola       Lower Middl~ Sub-s~  1972         0.729           7.23
##  7 Antigua and~ High Income  Latin~  1972         5.57            3.33
##  8 Argentina    Upper Middl~ Latin~  1972         3.64            3.15
##  9 Armenia      Upper Middl~ Europ~  1972        NA               3.03
## 10 Aruba        High Income  Latin~  1972        NA               2.69
## # … with 7,694 more rows, and 11 more variables: forest_area <dbl>,
## #   gdp_per_capita <dbl>, health_expenditure <dbl>,
## #   labor_force_participation <dbl>, life_expectancy <dbl>,
## #   malnutrition_prevalence <dbl>, population_total <dbl>,
## #   urban_population <dbl>, fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Ordenar filas según columnas

Ordenar según la columna *year* (descendente) e *income_group*

```
datosONU_tidy %>%
  arrange(-year, income_group)
```

```
## # A tibble: 7,704 x 17
##    country_name income_group region  year co2_emissions fertility_rate
##    <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
##  1 Andorra      High Income  Europ~  2007          6.63           1.18
##  2 Antigua and~ High Income  Latin~  2007          5.26           2.18
##  3 Aruba        High Income  Latin~  2007         23.3            1.74
##  4 Australia    High Income  East ~  2007         18.1            1.96
##  5 Austria      High Income  Europ~  2007          8.33           1.38
##  6 Bahamas, The High Income  Latin~  2007          4.52           1.88
##  7 Bahrain      High Income  Middl~  2007         21.7            2.29
##  8 Barbados     High Income  Latin~  2007          5.16           1.83
##  9 Belgium      High Income  Europ~  2007          9.71           1.82
## 10 Bermuda      High Income  North~  2007          7.97           1.76
## # ... with 7,694 more rows, and 11 more variables: forest_area <dbl>,
## #   gdp_per_capita <dbl>, health_expenditure <dbl>,
## #   labor_force_participation <dbl>, life_expectancy <dbl>,
## #   malnutrition_prevalence <dbl>, population_total <dbl>,
## #   urban_population <dbl>, fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Dejar valores únicos

Tantos valores como observaciones hay

```
datosONU_tidy %>%
  select(income_group)
```

```
## # A tibble: 7,704 x 1
##    income_group
##    <chr>
##  1 Low Income
##  2 Low Income
##  3 Low Income
##  4 Low Income
##  5 Low Income
##  6 Low Income
##  7 Low Income
##  8 Low Income
##  9 Low Income
## 10 Low Income
## # … with 7,694 more rows
```

Pero son pocos valores únicos/distintos

```
datosONU_tidy %>%
  distinct(income_group)
```

```
## # A tibble: 4 x 1
##   income_group
##   <chr>
## 1 Low Income
## 2 Upper Middle Income
## 3 High Income
## 4 Lower Middle Income
```

Dejar valores únicos

Se puede hacer para cualquier combinación de columnas/variables

```
datosONU_tidy %>%
  distinct(income_group, region) %>%
  arrange(income_group, region)
```

```
## # A tibble: 24 x 2
##    income_group region
##    <chr>        <chr>
##  1 High Income  East Asia and Pacific
##  2 High Income  Europe and Central Afica
##  3 High Income  Latin America and the Caribbean
##  4 High Income  Middle East and North Africa
##  5 High Income  North America
##  6 High Income  Sub-saharan Africa
##  7 Low Income   East Asia and Pacific
##  8 Low Income   Europe and Central Afica
##  9 Low Income   Latin America and the Caribbean
## 10 Low Income   Middle East and North Africa
## # ... with 14 more rows
```

Ejercicio

Ejercicio - Script

- `EjercicioManejo.R`

Manipular dos o más data frames

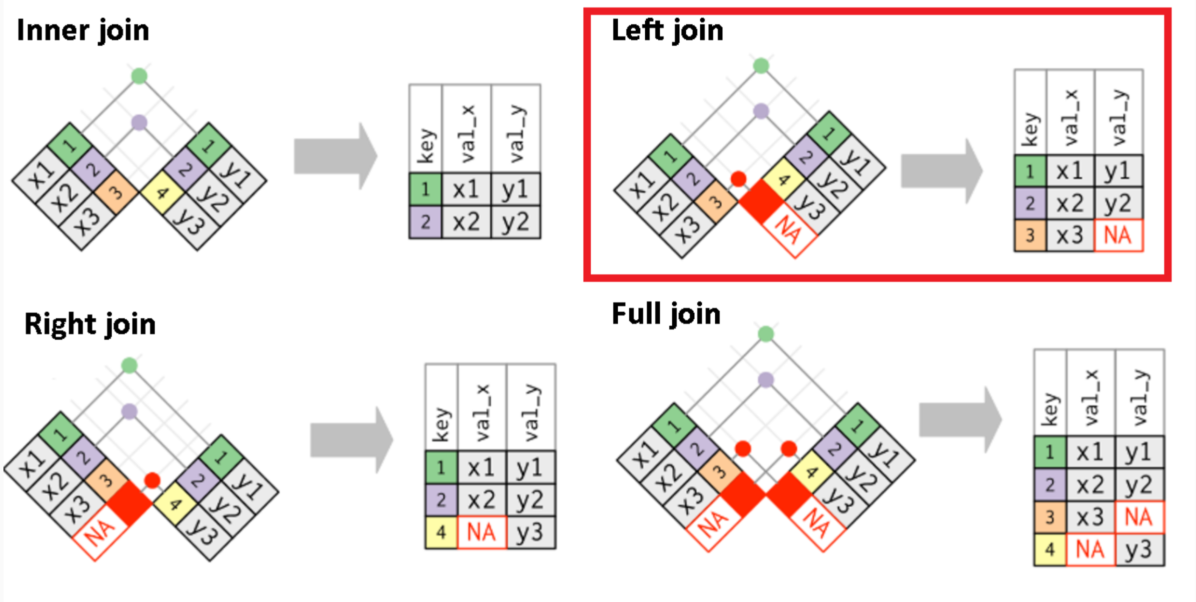
Trabajar con dos o más *data frames*

- *mutating joins*
  - `left_join`, `right_join`, `inner_join`, `full_join`
- *filtering joins*
  - `semi_join`, `anti_join`
- *set operations*
  - `intersect`, `union`, `setdiff`

Más información en https://dplyr.tidyverse.org/articles/two-table.html

Mutating joins

- Permiten combinar variables desde distintas tablas
- Generalmente el más utilizado es `left_join`



Inner join

key	val_x	val_y
1	x1	y1
2	x2	y2

Left join

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

Right join

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

Full join

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

Left Join

Digamos que queremos calcular el promedio de *fertility_rate* para cada *income_group* **pero nuestra tabla no tiene información sobre el grupo de ingresos**.

```
countries_noincomegroup
```

```
## # A tibble: 7,704 x 3
##    country_name  year fertility_rate
##    <chr>        <dbl>          <dbl>
##  1 Afghanistan   1972           7.67
##  2 Afghanistan   1973           7.67
##  3 Afghanistan   1974           7.67
##  4 Afghanistan   1975           7.67
##  5 Afghanistan   1976           7.67
##  6 Afghanistan   1977           7.67
##  7 Afghanistan   1978           7.67
##  8 Afghanistan   1979           7.67
##  9 Afghanistan   1980           7.67
## 10 Afghanistan   1981           7.67
## # … with 7,694 more rows
```

Left Join

Digamos que queremos calcular el promedio de *fertility_rate* para cada *income_group* pero nuestra tabla no tiene información sobre el grupo de ingresos. **Pero si tenemos otra tabla que asocia cada país a su grupo de ingresos**.

```
countries_noincomegroup
```

```
## # A tibble: 7,704 x 3
##    country_name  year fertility_rate
##    <chr>        <dbl>          <dbl>
##  1 Afghanistan   1972           7.67
##  2 Afghanistan   1973           7.67
##  3 Afghanistan   1974           7.67
##  4 Afghanistan   1975           7.67
##  5 Afghanistan   1976           7.67
##  6 Afghanistan   1977           7.67
##  7 Afghanistan   1978           7.67
##  8 Afghanistan   1979           7.67
##  9 Afghanistan   1980           7.67
## 10 Afghanistan   1981           7.67
## # … with 7,694 more rows
```

```
income_group
```

```
## # A tibble: 214 x 2
##    country_name        income_group
##    <chr>               <chr>
##  1 Afghanistan         Low Income
##  2 Albania             Upper Middle Income
##  3 Algeria             Upper Middle Income
##  4 American Samoa      Upper Middle Income
##  5 Andorra             High Income
##  6 Angola              Lower Middle Income
##  7 Antigua and Barbuda High Income
##  8 Argentina           Upper Middle Income
##  9 Armenia             Upper Middle Income
## 10 Aruba               High Income
## # … with 204 more rows
```

Left Join

```
countries_noincomegroup %>%
  left_join(income_group, by = "country_name")
```

```
## # A tibble: 7,704 x 4
##    country_name  year fertility_rate income_group
##    <chr>        <dbl>          <dbl> <chr>
##  1 Afghanistan   1972           7.67 Low Income
##  2 Afghanistan   1973           7.67 Low Income
##  3 Afghanistan   1974           7.67 Low Income
##  4 Afghanistan   1975           7.67 Low Income
##  5 Afghanistan   1976           7.67 Low Income
##  6 Afghanistan   1977           7.67 Low Income
##  7 Afghanistan   1978           7.67 Low Income
##  8 Afghanistan   1979           7.67 Low Income
##  9 Afghanistan   1980           7.67 Low Income
## 10 Afghanistan   1981           7.67 Low Income
## # … with 7,694 more rows
```

Left Join

¿Y si los nombres no son iguales?

```
names(income_group2)
```

```
## [1] "income_group" "country"
```

```
names(countries_noincomegroup)
```

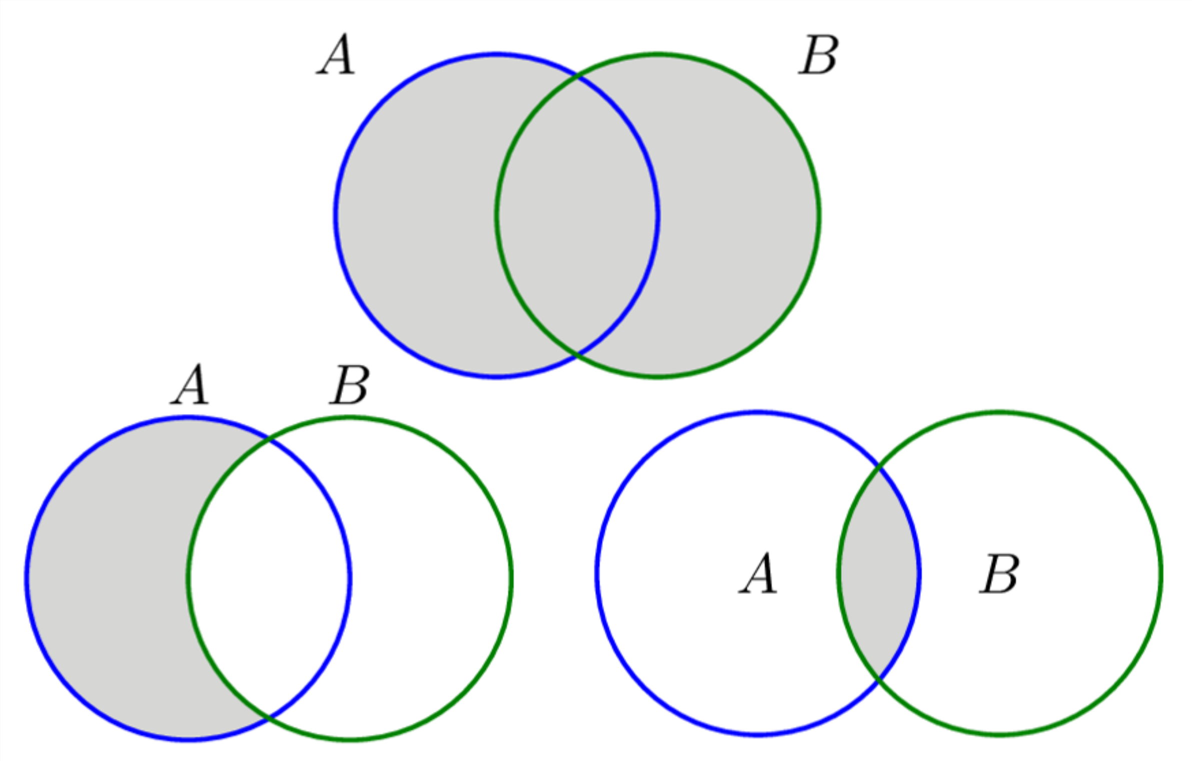
```
## [1] "country_name"   "year"           "fertility_rate"
```

```
countries_noincomegroup %>%
  left_join(income_group2, by = c("country_name" = "country"))
```

```
## # A tibble: 7,704 x 4
##    country_name  year fertility_rate income_group
##    <chr>        <dbl>          <dbl> <chr>
##  1 Afghanistan   1972           7.67 Low Income
##  2 Afghanistan   1973           7.67 Low Income
##  3 Afghanistan   1974           7.67 Low Income
##  4 Afghanistan   1975           7.67 Low Income
##  5 Afghanistan   1976           7.67 Low Income
##  6 Afghanistan   1977           7.67 Low Income
##  7 Afghanistan   1978           7.67 Low Income
##  8 Afghanistan   1979           7.67 Low Income
##  9 Afghanistan   1980           7.67 Low Income
## 10 Afghanistan   1981           7.67 Low Income
## # ... with 7,694 more rows
```

Set operations

Menos usadas en general pero útiles cuando se requieren



Set operations

Estas funciones esperan que `x` e `y` tengan las mismas variables/columnas y compara sus observaciones/filas

- `intersect(x, y)`: devuelve solo valores que estén presentes en `x` y en `y`
- `union(x, y)`: devuelve todos los valores (únicos) de `x` y de `y`
- `setdiff(x, y)`: devuelve observaciones que estén en `x` y no en `y`
    - `setdiff(y, x)`: devuelve observaciones estén en `y` y no en `x`

```r
df1 <- datosONU_tidy %>% slice(1:10)
df2 <- datosONU_tidy %>% slice(5:15)
```

Set operations

La intersección corresponde a las filas 5, 6, 7, 8, 9, y 10 de la base original

```
intersect(df1, df2)
```

```
## # A tibble: 6 x 17
##   country_name income_group region  year co2_emissions fertility_rate
##   <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
## 1 Afghanistan  Low Income   South~  1976         0.155           7.67
## 2 Afghanistan  Low Income   South~  1977         0.183           7.67
## 3 Afghanistan  Low Income   South~  1978         0.164           7.67
## 4 Afghanistan  Low Income   South~  1979         0.169           7.67
## 5 Afghanistan  Low Income   South~  1980         0.134           7.67
## 6 Afghanistan  Low Income   South~  1981         0.153           7.67
## # … with 11 more variables: forest_area <dbl>, gdp_per_capita <dbl>,
## #   health_expenditure <dbl>, labor_force_participation <dbl>,
## #   life_expectancy <dbl>, malnutrition_prevalence <dbl>,
## #   population_total <dbl>, urban_population <dbl>,
## #   fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Set operations

La unión corresponde a las primeras 15 filas de la base original

```
union(df1, df2)
```

```
## # A tibble: 15 x 17
##    country_name income_group region  year co2_emissions fertility_rate
##    <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
##  1 Afghanistan  Low Income   South~  1972         0.132           7.67
##  2 Afghanistan  Low Income   South~  1973         0.137           7.67
##  3 Afghanistan  Low Income   South~  1974         0.156           7.67
##  4 Afghanistan  Low Income   South~  1975         0.169           7.67
##  5 Afghanistan  Low Income   South~  1976         0.155           7.67
##  6 Afghanistan  Low Income   South~  1977         0.183           7.67
##  7 Afghanistan  Low Income   South~  1978         0.164           7.67
##  8 Afghanistan  Low Income   South~  1979         0.169           7.67
##  9 Afghanistan  Low Income   South~  1980         0.134           7.67
## 10 Afghanistan  Low Income   South~  1981         0.153           7.67
## 11 Afghanistan  Low Income   South~  1982         0.166           7.67
## 12 Afghanistan  Low Income   South~  1983         0.206           7.67
## 13 Afghanistan  Low Income   South~  1984         0.239           7.68
## 14 Afghanistan  Low Income   South~  1985         0.304           7.68
## 15 Afghanistan  Low Income   South~  1986         0.279           7.68
## # … with 11 more variables: forest_area <dbl>, gdp_per_capita <dbl>,
## #   health_expenditure <dbl>, labor_force_participation <dbl>,
## #   life_expectancy <dbl>, malnutrition_prevalence <dbl>,
## #   population_total <dbl>, urban_population <dbl>,
## #   fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Set operations

Las filas que están en `df1` y no en `df2` corresponden a la 1, 2, 3, y 4 de la base original

```
setdiff(df1, df2)
```

```
## # A tibble: 4 x 17
##   country_name income_group region  year co2_emissions fertility_rate
##   <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
## 1 Afghanistan  Low Income   South~  1972         0.132           7.67
## 2 Afghanistan  Low Income   South~  1973         0.137           7.67
## 3 Afghanistan  Low Income   South~  1974         0.156           7.67
## 4 Afghanistan  Low Income   South~  1975         0.169           7.67
## # … with 11 more variables: forest_area <dbl>, gdp_per_capita <dbl>,
## #   health_expenditure <dbl>, labor_force_participation <dbl>,
## #   life_expectancy <dbl>, malnutrition_prevalence <dbl>,
## #   population_total <dbl>, urban_population <dbl>,
## #   fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Set operations

Las filas que están en `df2` y no en `df1` corresponden a la 11, 12, 13, 14, y 15 de la base original

```
setdiff(df2, df1)
```

```
## # A tibble: 5 x 17
##   country_name income_group region  year co2_emissions fertility_rate
##   <chr>        <chr>        <chr>  <dbl>         <dbl>          <dbl>
## 1 Afghanistan  Low Income   South~  1982         0.166           7.67
## 2 Afghanistan  Low Income   South~  1983         0.206           7.67
## 3 Afghanistan  Low Income   South~  1984         0.239           7.68
## 4 Afghanistan  Low Income   South~  1985         0.304           7.68
## 5 Afghanistan  Low Income   South~  1986         0.279           7.68
## # … with 11 more variables: forest_area <dbl>, gdp_per_capita <dbl>,
## #   health_expenditure <dbl>, labor_force_participation <dbl>,
## #   life_expectancy <dbl>, malnutrition_prevalence <dbl>,
## #   population_total <dbl>, urban_population <dbl>,
## #   fossil_fuel_consumption <dbl>, poverty <dbl>,
## #   public_spending_education <dbl>
```

Ejercicio

Ejercicio - Script

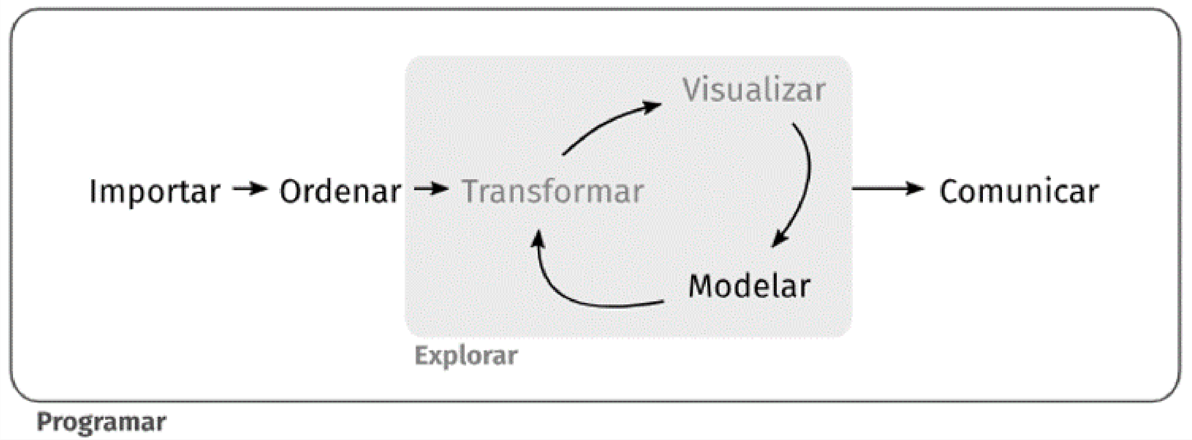
- `EjercicioDosTablas.R`

Para practicar

- `Ejercicios_dplyr_ggplot.pdf`
- `Ejercicios_dplyr_ggplot_RESPUESTAS.pdf`

Siguiente clase

- Más manejo de datos
- Transformación de datos
- Introducción a `R Markdown`


Importar
Ordenar
Transformar
Visualizar
Modelar
Comunicar
Explorar
Programar