

rparatodes (2)

 el maravilloso mundo de tidyverse 

Pau

06/09/2018

¿Qué es tidy?

```
install.packages('tidyverse') # solo se hace una vez!  
library(tidyverse)
```

3 reglas para datos :

- 1) cada variable tiene su propia columna
- 2) cada observación tiene su propia fila
- 3) cada valor tiene su propia celda

country	year	cases	population
Afghanistan	1999	1815	120000071
Afghanistan	2000	2666	20095360
Brazil	1999	30737	17206362
Brazil	2000	80488	17404898
China	1999	211258	1272015272
China	2000	210766	128002583

variables

country	year	cases	population
Afghanistan	1999	1815	120000071
Afghanistan	2000	2666	20095360
Brazil	1999	30737	17206362
Brazil	2000	80488	17404898
China	1999	211258	1272015272
China	2000	210766	128002583

observations

country	year	cases	population
Afghanistan	1999	1815	120000071
Afghanistan	2000	2666	20095360
Brazil	1999	30737	17206362
Brazil	2000	80488	17404898
China	1999	211258	1272015272
China	2000	210766	128002583

values

Concepto clave: %>%

El operador %>% funciona así:

$f(x)$ es igual a $x \%>\% f()$

Se lee como entonces o después permite leer de izquierda a derecha:

```
mis_datos %>%  
  hace_esta_cosa() %>%  
  ahora_esta_otra() %>%  
  y_una_mas()
```

```
resultado <- mis_datos %>%  
  hace_esta_cosa() %>%  
  ahora_esta_otra() %>%  
  y_una_mas()
```

datos ejemplo: planeta feliz

```
datos <- read.csv("http://www.lock5stat.com/datasets/HappyPlanetIndex.csv")
str(datos)
```

```
## 'data.frame':    143 obs. of  11 variables:
## $ Country       : Factor w/ 143 levels "Albania","Algeria",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ Region        : int  7 3 4 1 7 2 2 7 5 7 ...
## $ Happiness      : num  5.5 5.6 4.3 7.1 5 7.9 7.8 5.3 5.3 5.8 ...
## $ LifeExpectancy: num  76.2 71.7 41.7 74.8 71.7 80.9 79.4 67.1 63.1 68.7 ...
## $ Footprint      : num  2.2 1.7 0.9 2.5 1.4 7.8 5 2.2 0.6 3.9 ...
## $ HLY            : num  41.7 40.1 17.8 53.4 36.1 63.7 61.9 35.4 33.1 40.1 ...
## $ HPI            : num  47.9 51.2 26.8 59 48.3 ...
## $ HPIRank        : int  54 40 130 15 48 102 57 85 31 104 ...
## $ GDPperCapita   : int  5316 7062 2335 14280 4945 31794 33700 5016 2053 7918 ...
## $ HDI            : num  0.801 0.733 0.446 0.869 0.775 0.962 0.948 0.746 0.547 0.804 ...
## $ Population     : num  3.15 32.85 16.1 38.75 3.02 ...
```

Los datos tienen 11 variables:

- Region: 1 = Latin America, 2 = Western nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = former Communist countries
- Happiness Scored on a 0-10 scale for average level of happiness (10 is happiest)
- LifeExpectancy Average life expectancy (in years)
- Footprint Ecological footprint - a measure of the (per capita) ecological impact
- HLY Happy Life Years - combines life expectancy with well-being
- HPI Happy Planet Index (0-100 scale)
- HPIRank HPI rank for the country
- GDPperCapita Gross Domestic Product (per capita)
- HDI Human Development Index
- Population Population (in millions)

ordenando variables: arrange

```
mis_datos %>% arrange(variable)
```

Orden descendiente:

```
mis_datos %>% arrange(-variable)
```

```
mis_datos %>% arrange(desc(variable))
```

Para ordenar una variable y luego, la otra:

```
mis_datos %>% arrange(variable_1, variable_2)
```

ejemplo: arrange

```
asc <- datos %>%  
  arrange(Region)
```

##	Country	Region	Happiness	LifeExpectancy	Footprint	HLY	HPI	HPIRank
## 1	Argentina	1	7.1	74.8	2.5	53.4	58.95	15
## 2	Belize	1	6.6	75.9	2.6	50.2	54.53	27
## 3	Bolivia	1	6.5	64.7	2.1	42.1	49.35	47
## 4	Brazil	1	7.6	71.7	2.4	54.3	61.01	9
## 5	Chile	1	6.3	78.3	3.0	49.2	49.72	46
## 6	Colombia	1	7.3	72.3	1.8	53.0	66.10	6
##	GDPperCapita	HDI	Population					
## 1	14280	0.869	38.75					
## 2	7109	0.778	0.29					
## 3	2819	0.695	9.18					
## 4	8402	0.800	186.83					
## 5	12027	0.867	16.30					
## 6	7304	0.791	44.95					

```
desc <- datos %>%
  arrange(-Region)
```

##		Country	Region	Happiness	LifeExpectancy	Footprint	HLY
## 1		Albania	7	5.5	76.2	2.2	41.7
## 2		Armenia	7	5.0	71.7	1.4	36.1
## 3		Azerbaijan	7	5.3	67.1	2.2	35.4
## 4		Belarus	7	5.8	68.7	3.9	40.1
## 5	Bosnia and Herzegovina		7	5.9	74.5	2.9	44.0
## 6	Bulgaria		7	5.5	72.7	2.7	39.8
##	HPI	HPIRank	GDPperCapita	HDI	Population		
## 1	47.91	54	5316	0.801	3.15		
## 2	48.28	48	4945	0.775	3.02		
## 3	41.21	85	5016	0.746	8.39		
## 4	35.67	104	7918	0.804	9.78		
## 5	44.96	65	7032	0.803	3.78		
## 6	42.04	82	9032	0.824	7.74		

escogiendo variables: select

Se queda con todas las filas pero retiene solo algunas variables (columnas)

```
my_data %>%  
  select(VARIABLE1, VARIABLE2)
```

Elimina variables:

```
my_data %>%  
  select(-variable_1, -variable_2)
```

ejemplo: select

```
feliz_chico <- datos %>%  
  select(Country, Region, Happiness)
```

```
##      Country Region Happiness  
## 1   Albania      7      5.5  
## 2   Algeria      3      5.6  
## 3    Angola      4      4.3  
## 4 Argentina      1      7.1  
## 5  Armenia      7      5.0  
## 6 Australia      2      7.9
```

enfocándonos en ciertos casos: filter

```
mis_datos %>% filter(una_expresion_logica)
```

Que se cumplan dos condiciones:

```
mis_datos %>% filter(una_expresion_logica & otra_expresion_logica)
```

Que se cumpla una u otra condición:

```
mis_datos %>% filter(una_expresion_logica | otra_expresion_logica)
```

ejemplo: filter

```
feliz2 <- datos %>%  
  filter(Region == 2)
```

##	Country	Region	Happiness	LifeExpectancy	Footprint	HLY	HPI	HPIRank
## 1	Australia	2	7.9	80.9	7.8	63.7	36.64	102
## 2	Austria	2	7.8	79.4	5.0	61.9	47.69	57
## 3	Belgium	2	7.6	78.8	5.1	60.0	45.36	64
## 4	Canada	2	8.0	80.3	7.1	64.0	39.40	89
## 5	Cyprus	2	7.2	79.0	4.5	56.6	46.19	62
## 6	Denmark	2	8.1	77.9	8.0	62.9	35.47	105
##	GDPperCapita	HDI	Population					
## 1	31794	0.962	20.40					
## 2	33700	0.948	8.23					
## 3	32119	0.946	10.48					
## 4	33375	0.961	32.31					
## 5	22699	0.903	0.76					
## 6	33973	0.949	5.42					

```
feliz3 <- datos %>%
  filter(Happiness > 7)
```

```
##      Country Region Happiness LifeExpectancy Footprint  HLY   HPI HPIRank
## 1 Argentina      1        7.1          74.8        2.5 53.4 58.95      15
## 2 Australia      2        7.9          80.9        7.8 63.7 36.64     102
## 3  Austria      2        7.8          79.4        5.0 61.9 47.69      57
## 4  Belgium      2        7.6          78.8        5.1 60.0 45.36      64
## 5   Brazil      1        7.6          71.7        2.4 54.3 61.01       9
## 6   Canada      2        8.0          80.3        7.1 64.0 39.40      89
## GDPperCapita  HDI Population
## 1         14280 0.869        38.75
## 2         31794 0.962        20.40
## 3         33700 0.948         8.23
## 4         32119 0.946        10.48
## 5          8402 0.800       186.83
## 6         33375 0.961        32.31
```

arrange, filter & select

Recordatorio: arrange, filter & select no alteran el dataset original (mis_datos)

```
nuevos_datos <- viejos_datos %>%  
  filter(algunas_filas) %>%  
  select(algunas_columnas) %>%  
  arrange(por_variable)
```

Para alterar el dataset original:

```
viejos_datos <- viejos_datos %>%  
  filter(algunas_filas) %>%  
  select(algunas_columnas) %>%  
  arrange(por_variable)
```

creando nuevas variables: mutate

```
mis_datos <- mis_datos %>%  
  mutate(variable = expresión)
```

ejemplo: mutate

```
feliz <- datos %>%  
  mutate(TotalGDP = GDPperCapita * Population )
```

##	Country	Region	Happiness	LifeExpectancy	Footprint	HLY	HPI	HPIRank
## 1	Albania	7	5.5	76.2	2.2	41.7	47.91	54
## 2	Algeria	3	5.6	71.7	1.7	40.1	51.23	40
## 3	Angola	4	4.3	41.7	0.9	17.8	26.78	130
## 4	Argentina	1	7.1	74.8	2.5	53.4	58.95	15
## 5	Armenia	7	5.0	71.7	1.4	36.1	48.28	48
## 6	Australia	2	7.9	80.9	7.8	63.7	36.64	102

##	GDPperCapita	HDI	Population	TotalGDP
## 1	5316	0.801	3.15	16745.4
## 2	7062	0.733	32.85	231986.7
## 3	2335	0.446	16.10	37593.5
## 4	14280	0.869	38.75	553350.0
## 5	4945	0.775	3.02	14933.9
## 6	31794	0.962	20.40	648597.6

renombrando variables: rename

```
mis_datos <- mis_datos %>%  
  rename(nuevo_nombre = viejo_nombre)
```

ejemplo: rename

```
datos_es <- datos %>%  
  rename(pais = Country,  
         felicidad = Happiness)
```

```
##      pais Region felicidad LifeExpectancy Footprint  HLY   HPI HPIRank  
## 1  Albania     7      5.5          76.2      2.2 41.7 47.91     54  
## 2  Algeria     3      5.6          71.7      1.7 40.1 51.23     40  
## 3  Angola      4      4.3          41.7      0.9 17.8 26.78    130  
## 4 Argentina    1      7.1          74.8      2.5 53.4 58.95     15  
## 5  Armenia     7      5.0          71.7      1.4 36.1 48.28     48  
## 6 Australia    2      7.9          80.9      7.8 63.7 36.64    102  
## GDPperCapita  HDI Population  
## 1      5316 0.801      3.15  
## 2      7062 0.733     32.85  
## 3      2335 0.446     16.10  
## 4     14280 0.869     38.75  
## 5      4945 0.775      3.02  
## 6     31794 0.962     20.40
```

resúmenes agrupados: group & summarise

```
resumen <- mis_datos %>%  
  group_by(variable_para_agrupar) %>%  
  summarise(  
    mediana = median(variable),  
    media = mean(variable),  
    des_est = sd(variable))
```

resumen

```
resumen <- mis_datos %>%  
  group_by(variable_para_agrupar) %>%  
  summarise(  
    mediana = median(variable, na.rm = T),  
    media = mean(variable, na.rm = T),  
    des_est = sd(variable, na.rm = T))
```

resumen

```
resumen <- mis_datos %>%  
  group_by(variable_para_agrupar) %>%  
  summarise(cuenta = n())
```

```
resumen
```

ejemplo: group & summarise

```
resumen <- datos %>%  
  group_by(Region) %>%  
  summarise(AverageHappy = mean(Happiness))
```

```
## # A tibble: 6 x 2  
##   Region AverageHappy  
##   <int>         <dbl>  
## 1     1           6.91  
## 2     2           7.55  
## 3     3           5.99  
## 4     4           4.05  
## 5     5           5.59  
## 6     6           6.32
```

en resumen...

- 1) %>%: agiliza el flujo de trabajo
- 2) arrange: ordena variables
- 3) select: elige variables
- 4) filter: elige filas
- 5) mutate: crear nuevas variables
- 6) group_by and summarize: crea resúmenes agrupados