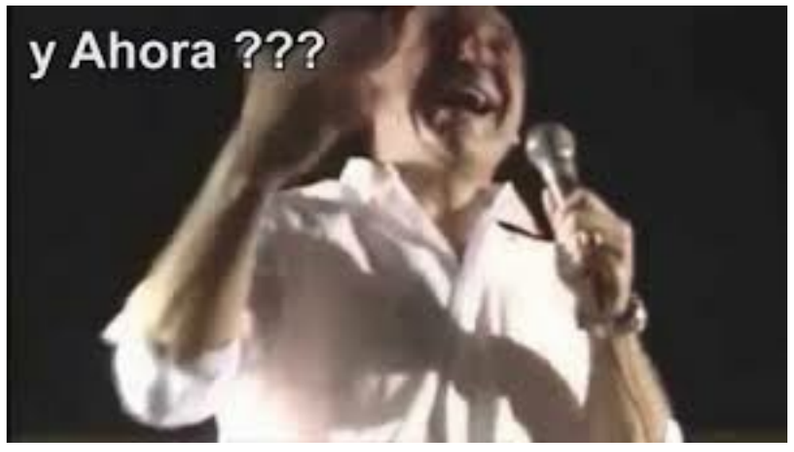
Group by y Summarise

Clases Spark R

Alex bajaña

2019-03-21



`group by` y `summarise` son herramientas para analizar los datos de en grupos ¡de cualquier tamaño!.

Ventajas:

- Similar al uso de `GROUP BY` en SQL.
- Mismas variantes que mutate

Empecemos

Funciones auxiliares

Tipo	Funciones	Descripción
Basicas	mean()	Media de un vector
Basicas	median()	Mediana de un vector
Basicas	sum()	Suma de elementos
Variaciones	sd()	Desviación standar
Variaciones	IQR()	Rango Interquantilico
Rangos	min()	Mínimo
Rangos	max()	Máximo
Rangos	quantile()	Cuantiles
Posición	first()	Primer elemento en el grupo
Posición	last()	Último elemento en el grupo
Posición	nth()	Elemento que se úbica en la posición n-sima
Conteo	n()	Número de filas
Conteo	n_distinct()	Número de elementos distintos

Datos

```
library(datasets)
library(tibble)

data <- datasets::Titanic

data <- data %>% as.tibble()

data %>% head(6) %>%  kable("html")
```

Class	Sex	Age	Survived	n
1st	Male	Child	No	0
2nd	Male	Child	No	0
3rd	Male	Child	No	35
Crew	Male	Child	No	0
1st	Female	Child	No	0
2nd	Female	Child	No	0

Variantes

summarise ¿con o sin group_by?

Con group_by

```
data %>%
group_by(Class) %>%
summarise(mean_personas = sum(n)) %>%
kable("html")
```

Class	mean_personas
1st	325
2nd	285
3rd	706
Crew	885

Variantes

summarise ¿con o sin group_by?

Sin group_by

```
data %>%
summarise(mean_personas = sum(n)) %>%
kable("html")
```

mean_personas
2201

Variantes

summarise_at

De forma similar que en los `mutate` se puede agregar variables llamando el vector de nombres de columas.

```
set.seed(12345)

data <- data %>%
  mutate(Inc =
           case_when(Class == "Crew" ~ sample(1:100,1,replace = T),
                     Class == "3rd" ~ sample(101:500,1,replace = T),
                     Class == "2nd" ~ sample(501:1000,1,replace = T),
                     TRUE ~ sample(1000:5000,1,replace = T)),
         Inc = Inc + rnorm(n = nrow(.),mean = 500,sd = 20),
         Inc = round(Inc,2))

data_ag <- data %>%
  group_by(Class,Survived) %>%
  summarise_at(.vars = c("n","Inc"),
               .funs = funs(mean))
```

Summarise_at

Resultado

Class	Survived	n	Inc
1st	No	30.50	5043.270
1st	Yes	50.75	5049.033
2nd	No	41.75	1390.525
2nd	Yes	29.50	1388.815
3rd	No	132.00	950.030
3rd	Yes	44.50	960.215
Crew	No	168.25	560.255
Crew	Yes	53.00	596.655

Se observa que el ingreso no difiere en promedio entre las distintas clases de tickets, asi mismo se ve que los no sobrevivientes al hundimiento del Titanic se encuentran en las clases tercera y personal.

Variantes

summarise_if

Evaluar un predicado y colapsar aquellas variables que cumplen la condicion

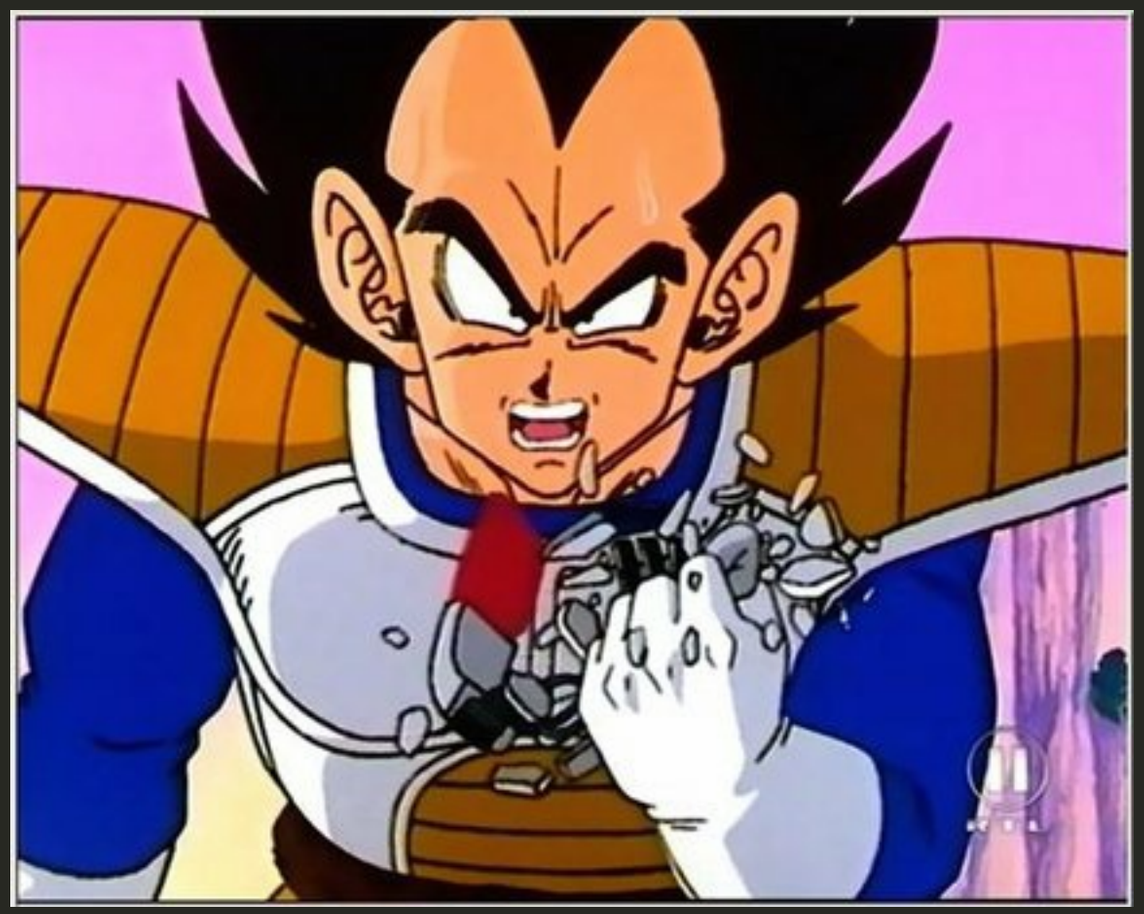
```
data_ag_2 <- data %>%
  group_by(Sex,Age) %>%
  summarise_if(is.numeric,funs(max = max(.),
                               min = min(.),
                               sd = sd(.),
                               Q25 = quantile(.,probs = 0.25),
                               Q50 = quantile(.,probs = 0.5),
                               Q75 = quantile(.,probs = 0.75)))
```

Summarise_if

Resultados

Sex	Age	n_max	n_min	n_sd	n_Q25	n_Q50	n_Q75
Female	Adult	140	3	50.303188	10.75	48.0	82.25
Female	Child	17	0	7.576986	0.00	0.5	13.25
Male	Adult	670	14	218.724182	70.50	136.0	240.75
Male	Child	35	0	12.118463	0.00	2.5	11.50

Sex	Age	Inc_max	Inc_min	Inc_sd	Inc_Q25	Inc_Q50	Inc_Q75
Female	Adult	5061.24	566.37	1904.860	851.3650	1194.660	2326.202
Female	Child	5057.60	554.61	1914.482	841.5625	1147.630	2289.637
Male	Adult	5042.68	569.75	1904.536	864.6575	1178.325	2321.855
Male	Child	5067.41	536.64	1915.204	872.8700	1169.260	2300.930



La similitud con mutate supera los 9000!

Ahora es su turno