Sesión 4.2 - Agrupamiento y resumen

July 7, 2021

#

Introducción al Análisis de Datos con R

0.1 Sesión 4.2: Agrupamiento y resumen

group_by() : Sirve para agrupar filas (filas) de un data frame para obtener variables de resumen.

summarize() : Sirve para obtener resumen de variables, dentro de esta podemos especifícar funciones resumen como : - sum() - mean() - sd() - var() - max() - min() - n() - n_dictinct() - first() - last()

pivot_longer()

pivot_wider()

[4]:

```r
library("datasets")
library("tidyverse")
```

[25]:

```r
# lectura de datos
iris <- read.csv('./datasets/iris.csv')
head(iris)
```

A data.frame: 6 × 5

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

[26]:

```r
# número de especies
unique(iris$Species)
```

1. 'setosa' 2. 'versicolor' 3. 'virginica'

0.2 ¿Cúanatas observaciones hay por especie?

[32]:

```
df01 <- iris %>%
    group_by(Species) %>%
    summarize('n'=n())
df01
```

A tibble: 3 × 2

Species	n
<chr>	<int>
setosa	50
versicolor	50
virginica	50

0.3 ¿Cúal es el promedio de cada medición por especie?

[34]:

```
df02 <- iris %>%
    group_by(Species) %>%
    summarize('Sepal_length'=mean(Sepal.Length),
              'Petal_length'=mean(Petal.Length),
              'Sepal_Width'=mean(Sepal.Width),
              'Petal_Width'=mean(Petal.Width))
```

[35]:

```
df02
```

A tibble: 3 × 5

Species	Sepal_length	Petal_length	Sepal_Width	Petal_Width
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
setosa	5.006	1.462	3.428	0.246
versicolor	5.936	4.260	2.770	1.326
virginica	6.588	5.552	2.974	2.026

0.4 ¿Cúal es la desviación estándar de cada medición por especie?

[36]:

```
df03 <- iris %>%
    group_by(Species) %>%
    summarize('Sepal_length'=sd(Sepal.Length),
              'Petal_length'=sd(Petal.Length),
              'Sepal_Width'=sd(Sepal.Width),
              'Petal_Width'=sd(Petal.Width))
```

[37]:

```
df03
```

A tibble: 3 × 5

Species	Sepal_length	Petal_length	Sepal_Width	Petal_Width
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
setosa	0.3524897	0.1736640	0.3790644	0.1053856
versicolor	0.5161711	0.4699110	0.3137983	0.1977527
virginica	0.6358796	0.5518947	0.3224966	0.2746501

1 Pivot tables

[13]:

```
url <- 'https://raw.githubusercontent.com/JulioCesarMartinez-00/
↪S1-Introduccion-a-R/main/Sesi%C3%B3n4%20Manejo%20de%20datos%20con%20R/
↪datsets/df_DownJones.csv'
```

[15]:

```
down <- read.csv(url) %>%
    mutate(Date = as.Date(Date, "%Y-%m-%d"))

head(down)
```

A data.frame: 6 × 27

	Date <date>	WMT <dbl>	MRK <dbl>	INTC <dbl>	MSFT <dbl>	MMM <dbl>	AAPL <dbl>	VZ <dbl>
1	1990-01-03	5.890625	13.31250	1.093750	0.619792	20.34375	0.334821	25.5744
2	1990-01-04	5.859375	13.10417	1.117188	0.638021	20.50000	0.335938	24.5908
3	1990-01-05	5.796875	12.83333	1.109375	0.622396	20.15625	0.337054	24.0287
4	1990-01-08	5.875000	13.00000	1.125000	0.631944	20.68750	0.339286	24.2255
5	1990-01-09	5.718750	12.89583	1.156250	0.630208	20.68750	0.335938	23.5510
6	1990-01-10	5.718750	12.72917	1.125000	0.612847	20.56250	0.321429	22.9327

1.1 Obtener el promedio del precio por mes y año de cada activo

[68]:

```
down_longer <- down %>%
    mutate(year = substring(Date,1,4),
           month = substring(Date,6,7),
           day = substring(Date, 9,10)) %>%
    select(-Date) %>%
    pivot_longer(cols=WMT:BA, names_to='name', values_to='price')
```

[69]:

```
head(down_longer)
```

A tibble: 6 × 5

year <chr>	month <chr>	day <chr>	name <chr>	price <dbl>
1990	01	03	WMT	5.890625
1990	01	03	MRK	13.312500
1990	01	03	INTC	1.093750
1990	01	03	MSFT	0.619792
1990	01	03	MMM	20.343750
1990	01	03	AAPL	0.334821

[74]:

```
down_avg <- down %>%
    mutate(year = substring(Date,1,4),
           month = substring(Date,6,7),
           day = substring(Date, 9,10)) %>%
    select(-Date) %>%
    pivot_longer(cols=WMT:BA, names_to='name', values_to='price') %>%
    group_by(year, month, name) %>%
    summarize(mean_price = mean(price)) %>%
    pivot_wider(names_from=name, values_from=mean_price)
```

[75]:

```
down_avg
```

A grouped_df: 373 × 28

year	month	AAPL	AXP	BA	CAT	CVX	DIS
<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1990	01	0.3110652	8.262480	20.07143	7.091518	16.43750	9.035087
1990	02	0.3026903	7.467038	20.73684	7.070724	17.05757	8.842264
1990	03	0.3447646	7.092516	23.00947	7.672585	17.01847	9.277864
1990	04	0.3646765	7.027706	23.83542	7.586719	16.58594	9.251520
1990	05	0.3666296	7.377797	25.81818	8.208097	17.16477	9.475912
1990	06	0.3612352	7.919168	29.02679	7.871280	17.55208	10.559856
1990	07	0.3916880	7.899244	29.67560	6.504464	18.62500	10.476669
1990	08	0.3427311	6.577039	25.04620	5.670516	19.37092	8.609536
1990	09	0.2944665	5.873009	22.43421	5.320724	19.17270	7.801679
1990	10	0.2583462	5.057323	22.66848	5.144701	17.61277	7.617227
1990	11	0.3152636	5.249303	22.46726	5.178571	17.24702	7.835726
1990	12	0.3718751	5.518206	22.76562	5.707031	17.65625	8.421730
1991	01	0.4377030	5.238923	23.22443	5.738636	18.00568	8.201683
1991	02	0.5220277	6.396447	24.54605	6.305099	18.13651	9.786037
1991	03	0.5818639	7.011613	24.18750	6.557031	19.15156	10.071035
1991	04	0.5783280	7.151035	23.61364	6.017756	19.31392	9.677229
1991	05	0.4322240	6.280563	23.38636	6.190341	18.79261	9.853790
1991	06	0.3912389	6.178009	23.81563	6.375000	17.85469	9.435975
1991	07	0.4037643	6.055265	22.43466	6.080256	17.84375	9.846316
1991	08	0.4643364	6.770661	23.61932	5.989347	17.70028	9.765042
1991	09	0.4487165	6.646308	25.14375	5.730469	18.02812	9.347601
1991	10	0.4544351	5.412763	24.60870	5.817255	18.81522	9.555825
1991	11	0.4590400	4.872886	23.46875	5.573437	17.66406	9.082479
1991	12	0.4587585	4.921318	21.94048	5.125744	16.74405	8.908839
1992	01	0.5589996	5.628075	25.24148	5.674716	16.76989	10.656255
1992	02	0.5828244	5.529132	23.73355	6.226974	15.70888	12.071645
1992	03	0.5652901	5.677817	22.50000	6.186790	15.73295	12.413926
1992	04	0.5141902	5.807186	22.59524	6.566964	16.59524	12.490284
1992	05	0.5406809	5.815922	21.87187	7.141406	17.20781	12.475628
1992	06	0.4498782	6.069894	21.35227	7.060369	17.61506	12.039784
...	...	...	...	...	...	...	...
2018	08	53.33652	103.48870	346.8278	138.5100	120.92478	112.9287
2018	09	55.51842	108.36263	358.5921	147.7458	118.68316	111.4163
2018	10	55.21141	104.51696	366.3756	136.6257	118.16913	115.2522
2018	11	47.80893	107.90714	343.6033	127.4086	117.17905	115.5005
2018	12	41.06658	101.14895	322.2274	125.1900	111.55053	109.8268
2019	01	38.54155	99.13143	352.3071	130.5943	112.24095	110.8986
2019	02	42.93197	106.03053	414.5958	134.4847	119.07211	112.1663
2019	03	45.82345	110.18191	390.7376	133.5600	123.68762	112.2843
2019	04	50.12905	112.44952	379.6605	139.9086	122.13619	125.8138
2019	05	47.81841	118.12909	355.7386	127.6164	118.99954	133.7032
2019	06	48.24225	121.96950	358.6600	128.9510	121.74600	138.5810
2019	07	51.30409	126.60000	357.4545	135.2345	124.61364	142.8732
2019	08	51.23943	122.33773	341.0827	118.3159	118.48409	136.5409
2019	09	54.49875	118.58200	375.4620	127.6480	121.22700	135.1235
2019	10	58.82163	116.32478	359.1278	129.5178	115.63435	130.1404
2019	11	65.63013	120.06850	363.7250	145.2240	119.51000	142.8855
2019	12	69.13143	122.48238	339.1376	145.1448	118.49571	146.7581
2020	01	77.97905	129.24476	325.8671	143.6243	115.09000	142.8595
2020	02	77.81763	129.70000	326.0979	133.8768	106.70263	136.6874
2020	03	65.61102	92.51500	178.7441	107.2591	76.46409	101.8877