How can I calculate lag by group using dplyr?

Authored by

stats writer

July 1, 2024

The process of calculating lag by group using dplyr involves using the grouping function in dplyr to separate the data into smaller subsets and then applying the lag function to each group. This allows for the calculation of lag values specific to each group, providing a comprehensive analysis of the data. By utilizing dplyr's functions, the process of calculating lag by group can be efficiently and accurately performed, making it a valuable tool for data analysis.

Calculate Lag by Group Using dplyr

You can use the following syntax to calculate lagged values by group in R using the package:

```
df %>%
group_by(var1) %>%
mutate(lag1_value = lag(var2, n=1, order_by=var1))
```

Note: The function adds a new variable to the data frame that contains the lagged values.

The following example shows how to use this syntax in practice.

Example: Calculate Lagged Values by Group Using dplyr

Suppose we have the following data frame in R that shows the sales made by two different stores during various days:

```
#create data frame
```

```
df <- data.frame(store=c('A', 'B', 'A', 'B', 'A', 'B', 'A', 'B'),
sales=c(7, 12, 10, 9, 9, 11, 18, 23))

#view data frame
df

store sales
1 A 7
2 B 12
3 A 10
4 B 9
5 A 9
6 B 11
7 A 18
8 B 23
```

We can use the following code to create a new column that shows the lagged values of sales for each store:

```
library(dplyr)

#calculate lagged sales by group
df %>%
group_by(store) %>%
mutate(lag1_sales = lag(sales, n=1, order_by=store))
```

 # A tibble: 8 x 3

 # Groups: store

store sales lag1_sales

1 A 7 NA

2 B 12 NA

3 A 10 7

4 B 9 12

5 A 9 10

6 B 11 9

7 A 18 9

8 B 23 11

Here's how to interpret the output:

The first value of lag1_sales is NA because there is no previous value for sales for store A.The second value of lag1_sales is NA because there is no previous value for sales for store B.The third value of lag1_sales is 7 because this is the previous value for sales for store A.The fourth value of lag1_sales is 12 because this is the previous value for sales for store B.

And so on.

Note that you can also change the number of lags used

by modifying the value for n in the lag() function.

Additional Resources

The following tutorials explain how to perform other common calculations in R: