How can I calculate a cumulative sum using dplyr?

Authored by

stats writer

July 1, 2024

Dplyr is a popular R package that offers a set of tools for data manipulation and analysis. One of its functions is the ability to calculate a cumulative sum, which is the sum of a series of numbers as they accumulate over time. To do this, the dplyr package provides the "cumsum()" function, which can be applied to a column or vector of numbers using the "mutate()" verb. This allows for efficient and concise computation of cumulative sums, making it a valuable tool in data analysis and modeling. Additionally, dplyr allows for easy integration with other functions and capabilities of the package, making it a powerful and versatile tool for calculating cumulative sums.

Calculate a Cumulative Sum Using dplyr

You can use the following methods to calculate the cumulative sum of a column in R using the package:

Method 1: Calculate Cumulative Sum of One Column

df %>% mutate(cum_sum = cumsum(var1))

Method 2: Calculate Cumulative Sum by Group

df %>% group_by(var1) %>% mutate(cum_sum = cumsum(var2))

The following examples show how to use each method in practice.

Example 1: Calculate Cumulative Sum Using dplyr

Suppose we have the following data frame in R:

```
#create dataset
df <- data.frame(day=c(1, 2, 3, 4, 5, 6, 7, 8),
sales=c(7, 12, 10, 9, 9, 11, 18, 23))

#view dataset
df

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

We can use the following code to create a new column that contains the cumulative sum of the values in the 'sales' column:

```
library(dplyr)

#calculate cumulative sum of sales
df %>% mutate(cum_sales = cumsum(sales))
```

day sales cum_sales

1 1 7 7

2 2 12 19

3 3 10 29

4 4 9 38

5 5 9 47

6 6 11 58

7 7 18 76

8 8 23 99

Example 2: Calculate Cumulative Sum by Group Using dplyr

Suppose we have the following data frame in R:

```
#create dataset
df <- data.frame(store=c('A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
day=c(1, 2, 3, 4, 1, 2, 3, 4),
sales=c(7, 12, 10, 9, 9, 11, 18, 23))

#view dataset
df
```

store day sales

1 A 1 7

2 A 2 12

3 A 3 10

4 A 4 9

5 B 1 9

6 B 2 11

7 B 3 18

8 B 4 23

We can use the following code to create a new column that contains the cumulative sum of the values in the 'sales' column, grouped by the 'store' column:

```
library(dplyr)

#calculate cumulative sum of sales by store
df %>% group_by(store) %>% mutate(cum_sales = cumsum(sales))
```

A tibble: 8 x 4

Groups: store

store day sales cum_sales

1 A 1 7 7

2 A 2 12 19

3 A 3 10 29

4 A 4 9 38

5 B 1 9 9

6 B 2 11 20

7 B 3 18 38

8 B 4 23 61

Additional Resources