How can you group data and count occurrences while also applying a condition using the R programming language?

Authored by

stats writer

June 27, 2024

In the R programming language, data can be grouped and occurrences can be counted while also applying a condition through the use of the "count" function. This function allows for the creation of a summary table that displays the number of occurrences for each unique value in a specified column or group of columns, while also allowing for the inclusion of a condition to filter the data. This method of grouping, counting, and applying a condition is useful for analyzing large datasets and identifying patterns or trends within the data. Additionally, the use of the R programming language allows for efficient and precise execution of these tasks.

R: Group By and Count with Condition

You can use the following basic syntax to perform a group by and count with condition in R:

```
library(dplyr)

df %>%
group_by(var1) %>%
summarize(count = sum(var2 == 'val'))
```

This particular syntax groups the rows of the data frame based on var1 and then counts the number of rows where var2 is equal to 'val.'

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

Example: Group By and Count with Condition in R

Suppose we have the following data frame in R that

contains information about various basketball players:

```
#create data frame
df <- data.frame(team=c('A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
pos=c('Gu', 'Fo', 'Fo', 'Fo', 'Gu', 'Gu', 'Fo', 'Fo'),
points=c(18, 22, 19, 14, 14, 11, 20, 28))

#view data frame
df

team pos points
1 A Gu 18
2 A Fo 22
3 A Fo 19
4 A Fo 14
5 B Gu 14
6 B Gu 11
7 B Fo 20
8 B Fo 28
```

The following code shows how to group the data frame by the team variable and count the number of rows where the pos variable is equal to 'Gu':

```
library(dplyr)
```

```
#group by team and count rows where pos is 'Gu'
df %>%
group_by(team) %>%
summarize(count = sum(pos == 'Gu'))

# A tibble: 2 x 2
team count

1 A 1
2 B 2
```

From the output we can see:

Team A has 1 row where the pos column is equal to 'Gu'Team B has 2 rows where the pos column is equal to 'Gu'

We can use similar syntax to perform a group by and count with some numerical condition.

For example, the following code shows how to group by the team variable and count the number of rows where the points variable is greater than 15:

```
library(dplyr)
```

```
#group by team and count rows where pos is 'Gu'
df %>%
group_by(team) %>%
summarize(count = sum(points > 15))

# A tibble: 2 x 2
team count

1 A 3
2 B 2
```

From the output we can see:

Team A has 3 rows where the points column is greater than 15Team B has 2 rows where the points column is greater than 15

You can use similar syntax to perform a group by and count with any specific condition you'd like.