“How can I use dplyr to summarise multiple columns simultaneously?”

Authored by

stats writer

June 27, 2024

Dplyr is a popular R package that allows users to efficiently manipulate and summarize data frames. With dplyr, it is possible to summarize multiple columns simultaneously by using the summarise_all function. This function allows for the application of a chosen summarizing function, such as mean or sum, across multiple columns at once. This provides a convenient and concise way to obtain summary statistics for multiple variables in a data frame. Additionally, dplyr's intuitive syntax makes it easy to specify the columns and summarizing function desired, making it a valuable tool for data analysis and summary.

Summarise Multiple Columns Using dplyr

You can use the following methods to summarise multiple columns in a data frame using dplyr:

Method 1: Summarise All Columns

```
#summarise mean of all columns
df %>%
group_by(group_var) %>%
summarise(across(everything(), mean, na.rm=TRUE))
```

Method 2: Summarise Specific Columns

```
#summarise mean of col1 and col2 only
df %>%
group_by(group_var) %>%
summarise(across(c(col1, col2), mean, na.rm=TRUE))
```

Method 3: Summarise All Numeric Columns

```
#summarise mean and standard deviation of all numeric columns
df %>%
group_by(group_var) %>%
summarise(across(where(is.numeric),  list(mean=mean, sd=sd), na.rm=TRUE))
```

The following examples show how to each method with the following data frame:

```
#create data frame
df <- data.frame(team=c('A', 'A', 'A', 'B', 'B', 'B'),
points=c(99, 90, 86, 88, 95, 90),
assists=c(33, 28, 31, 39, 34, 25),
rebounds=c(NA, 28, 24, 24, 28, 19))

#view data frame
df

team points assists rebounds
1 A 99 33 NA
2 A 90 28 28
3 A 86 31 24
4 B 88 39 24
5 B 95 34 28
```

6 B 90 25 19

Example 1: Summarise All Columns

The following code shows how to summarise the mean of all columns:

```
library(dplyr)

#summarise mean of all columns, grouped by team
df %>%
group_by(team) %>%
summarise(across(everything(), mean, na.rm=TRUE))

# A tibble: 2 x 4
team points assists rebounds

1 A 91.7 30.7 26
2 B 91 32.7 23.7
```

Example 2: Summarise Specific Columns

The following code shows how to summarise the mean of only the points and rebounds columns:

```
library(dplyr)
```

```
#summarise mean of points and rebounds, grouped by team
df %>%
group_by(team) %>%
summarise(across(c(points, rebounds), mean, na.rm=TRUE))

# A tibble: 2 x 3
team points rebounds

1 A 91.7 26
2 B 91 23.7
```

Example 3: Summarise All Numeric Columns

The following code shows how to summarise the mean and standard deviation for all numeric columns in the data frame:

```
library(dplyr)

#summarise mean and standard deviation of all numeric columns
df %>%
group_by(team) %>%
summarise(across(where(is.numeric),  list(mean=mean,
```

sd=sd), na.rm=TRUE))

A tibble: 2 x 7
team points_mean points_sd assists_mean assists_sd rebounds_mean rebounds_sd

1 A 91.7 6.66 30.7 2.52 26 2.83
2 B 91 3.61 32.7 7.09 23.7 4.51

The output displays the mean and standard deviation for all numeric variables in the data frame.

Note that in this example we used the list() function to list out several summary statistics that we wanted to calculate.

Note: In each example, we utilized the dplyr across() function. You can find the complete documentation for this function .

Additional Resources

The following tutorials explain how to perform other common functions using dplyr: