How can we calculate summary statistics by group in R?

Authored by

stats writer

May 12, 2024

Calculating summary statistics by group in R refers to a statistical method that allows users to calculate descriptive statistics for different subsets of a dataset based on a particular grouping variable. This can be achieved by using functions such as "aggregate" or "tapply" to group the data and then apply summary statistics, such as mean, median, standard deviation, etc., to each group. This method is useful for obtaining insights and comparing the characteristics of different groups within a dataset, providing a deeper understanding of the overall data. It is commonly used in data analysis and can be easily implemented in R programming language, making it a valuable tool for statistical analysis.

Calculate Summary Statistics by Group in R

There are two basic ways to calculate summary statistics by group in R:

Method 1: Use tapply() from Base R

tapply(df$value_col, df$group_col, summary)

Method 2: Use group_by() from dplyr Package

library(dplyr)

df %>%
group_by(group_col) %>%
summarize(min = min(value_col),
q1 = quantile(value_col, 0.25),
median = median(value_col),
mean = mean(value_col),

q3 = quantile(value_col, 0.75),
max = max(value_col))

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

Method 1: Use tapply() from Base R

The following code shows how to use the tapply() function from base R to calculate summary statistics by group:

```
#create data frame
df <- data.frame(team=c('A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
points=c(99, 68, 86, 88, 95, 74, 78, 93),
assists=c(22, 28, 31, 35, 34, 45, 28, 31),
rebounds=c(30, 28, 24, 24, 30, 36, 30, 29))

#calculate summary statistics of 'points' grouped by 'team'
tapply(df$points, df$team, summary)

$A
Min. 1st Qu. Median Mean 3rd Qu. Max.
68.00 81.50 87.00 85.25 90.75 99.00
```

```
$B
Min. 1st Qu. Median Mean 3rd Qu. Max.
74.0 77.0 85.5 85.0 93.5 95.0
```

Method 2: Use group_by() from dplyr Package

The following code shows how to use the group_by() and summarize() functions from the package to calculate summary statistics by group:

```
library(dplyr)

#create data frame
df <- data.frame(team=c('A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
points=c(99, 68, 86, 88, 95, 74, 78, 93),
assists=c(22, 28, 31, 35, 34, 45, 28, 31),
rebounds=c(30, 28, 24, 24, 30, 36, 30, 29))

#calculate summary statistics of 'points' grouped by 'team'
df %>%
group_by(team) %>%
summarize(min = min(points),
q1 = quantile(points, 0.25),
median = median(points),
```

mean = mean(points),

q3 = quantile(points, 0.75),

max = max(points))

 # A tibble: 2 x 7

team min q1 median mean q3 max

1 A 68 81.5 87 85.2 90.8 99

2 B 74 77 85.5 85 93.5 95

Notice that both methods return the exact same results.

It's worth noting that the dplyr approach will likely be faster for large data frames but both methods will perform similarly on smaller data frames.

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