How can we calculate summary statistics in R using dplyr?

Authored by

stats writer

June 25, 2024

Dplyr is a popular R package used for data manipulation and analysis. It provides a set of functions that can be used to calculate summary statistics in R. To calculate summary statistics, one must first import their data into R and load the dplyr package. The group_by() function can then be used to group the data by a specific variable. The summarise() function can then be used to calculate various summary statistics such as mean, median, standard deviation, and more. Additionally, dplyr also allows for the use of filters and pipes to further refine the data and calculate summary statistics for specific subsets. Overall, dplyr provides a convenient and efficient way to calculate summary statistics in R.

Calculate Summary Statistics in R Using dplyr

You can use the following syntax to calculate summary statistics for all numeric variables in a data frame in R using functions from the dplyr package:

```
library(dplyr)
library(tidyr)

df %>% summarise(across(where(is.numeric), .fns =
list(min = min,
median = median,
mean = mean,
stdev = sd,
q25 = ~quantile(., 0.25),
q75 = ~quantile(., 0.75),
max = max))) %>%
pivot_longer(everything(), names_sep='_',
names_to=c('variable', '.value'))
```

The summarise() function comes from the dplyr package and is used to calculate summary statistics for variables.

The pivot_longer() function comes from the tidyr package and is used to format the output to make it easier to read.

This particular syntax calculates the following summary statistics for each numeric variable in a data frame:

Minimum valueMedian valueMean valueStandard deviation25th percentile75th percentileMaximum value

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

Example: Calculate Summary Statistics in R Using dplyr

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'),
points=c(12, 15, 19, 14, 24, 25, 39, 34),
assists=c(6, 8, 8, 9, 12, 6, 8, 10),
```

rebounds=c(9, 9, 8, 10, 8, 4, 3, 3))

#view data frame
df

team points assists rebounds
1 A 12 6 9
2 A 15 8 9
3 A 19 8 8
4 A 14 9 10
5 B 24 12 8
6 B 25 6 4
7 B 39 8 3
8 B 34 10 3

We can use the following syntax to calculate summary statistics for each numeric variable in the data frame:

```
library(dplyr)
library(tidyr)

#calculate summary statistics for each numeric variable in data frame
df %>% summarise(across(where(is.numeric), .fns =
list(min = min,
```

```
median = median,
mean = mean,
stdev = sd,
q25 = ~quantile(., 0.25),
q75 = ~quantile(., 0.75),
max = max))) %>%
pivot_longer(everything(), names_sep='_',
names_to=c('variable', '.value'))
```

```
# A tibble: 3 x 8
variable min median mean stdev q25 q75 max

1 points 12 21.5 22.8 9.74 14.8 27.2 39
2 assists 6 8 8.38 2.00 7.5 9.25 12
3 rebounds 3 8 6.75 2.92 3.75 9 10
```

From the output we can see:

The minimum value in the points column is 12.The median value in the points column is 21.5.The mean value in the points column is 22.8.

And so on.

Note: In this example, we utilized the dplyr across()

function. You can find the complete documentation for this function .