How can data frames be merged based on multiple columns using the R programming language?

Authored by

stats writer

May 5, 2024

Data frames are a data structure commonly used in the R programming language to store and manipulate data. They consist of rows and columns, with each column representing a variable and each row representing an observation. In certain cases, it may be necessary to merge or combine data frames based on multiple columns. This can be achieved using the merge() function in R, which allows for merging data frames based on one or more columns that serve as common identifiers. By specifying the columns to merge on, this function combines the data from the two data frames into a single, merged data frame. This process is commonly used in data analysis and can help to combine multiple sources of data for a more comprehensive analysis.

R: Merge Data Frames Based on Multiple Columns

You can use the following basic syntax to merge two data frames in R based on multiple columns:

merge(df1, df2, by.x=c('col1', 'col2'), by.y=c('col1', 'col2'))

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

Example: Merge Data Frames on Multiple Columns

Suppose we have the following two data frames in R:

```
#define data frames
df1 = data.frame(playerID=c(1, 2, 3, 4, 5, 6),
team=c('A', 'B', 'B', 'B', 'C', 'C'),
points=c(19, 22, 25, 29, 34, 39))
```

```
df2 = data.frame(playerID=c(1, 2, 3, 4),
tm=c('A', 'B', 'B', 'B'),
rebounds=c(7, 8, 8, 14))

#view first data frame
df1
```

playerID team points
1 1 A 19
2 2 B 22
3 3 B 25
4 4 B 29
5 5 C 34
6 6 C 39

```
#view second data frame
df2
```

playerID tm rebounds
1 1 A 7
2 2 B 8
3 3 B 8
4 4 B 14

Notice that the two data frames share the playerID

column, but the team columns have different names in each data frame:

The first data frame has column 'team'The second data frame has column 'tm'

In order to merge these data frames based on the playerID and the team columns, we need to use the by.x and by.y arguments.

We can use the following code to perform this merge:

#merge two data framesmerged = merge(df1, df2, by.x=c('playerID', 'team'), by.y=c('playerID', 'tm'))

#view merged data frame
merged

playerID team points rebounds
1 1 A 19 7
2 2 B 22 8
3 3 B 25 8
4 4 B 29 14

The final merged data frame contains data for the four players that belong to both original data frames.

The following tutorials explain how to perform other common functions related to data frames in R: