

# How can I join two data frames on multiple columns using dplyr in R?

Authored by  
**stats writer**

June 29, 2024

## RECOMMENDED CITATION

stats writer (2024). *How can I join two data frames on multiple columns using dplyr in R?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=158908>

To join two data frames on multiple columns using dplyr in R, one can use the "inner\_join" function which merges two data frames based on their shared columns. This function allows for specifying multiple columns to join on, ensuring that the resulting data frame only includes rows where all specified columns have matching values. This process allows for efficient and accurate merging of data frames in R, which is helpful for data analysis and manipulation tasks.

## Join Data Frames on Multiple Columns Using dplyr

You can use the following basic syntax to join data frames in R based on multiple columns using dplyr:

```
library(dplyr)
```

```
left_join(df1, df2, by=c('x1'='x2', 'y1'='y2'))
```

This particular syntax will perform a left join where the following conditions are true:

The value in the x1 column of df1 matches the value in the x2 column of df2. The value in the y1 column of df1 matches the value in the y2 column of df2.

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

**Example: Join on Multiple Columns Using dplyr**

**Suppose we have the following two data frames in R:**

```
#define first data frame
```

```
df1 = data.frame(team=c('A', 'A', 'B', 'B'),  
pos=c('G', 'F', 'F', 'G'),  
points=c(18, 22, 19, 14))
```

```
df1
```

```
team pos points
```

```
1 A G 18
```

```
2 A F 22
```

```
3 B F 19
```

```
4 B G 14
```

```
#define second data frame
```

```
df2 = data.frame(team_name=c('A', 'A', 'B', 'C', 'C'),  
position=c('G', 'F', 'F', 'G', 'F'),  
assists=c(4, 9, 8, 6, 5))
```

```
df2
```

```
team_name position assists
```

```
1 A G 4
```

```
2 A F 9
```

```
3 B F 8
```

```
4 C G 6
```

## 5 C F 5

We can use the following syntax in dplyr to perform a left join based on two columns:

```
library(dplyr)
```

```
#perform left join based on multiple columns
```

```
df3 <- left_join(df1, df2, by=c('team'='team_name',  
'pos'='position'))
```

```
#view result
```

```
df3
```

```
team pos points assists
```

```
1 A G 18 4
```

```
2 A F 22 9
```

```
3 B F 19 8
```

```
4 B G 14 NA
```

The resulting data frame contains all rows from df1 and only the rows in df2 where the team and position values matched.

Also note that if the two data frames share the same

**column names, you can simply use the following syntax to join on multiple columns:**

```
library(dplyr)
```

```
#perform left join based on multiple columns
```

```
df3 <- left_join(df1, df2, by=c('team', 'position'))
```

**Additional Resources**

ARABPSYCHOLOGY.COM