How can I perform a left join using selected columns in dplyr?

Authored by

stats writer

June 23, 2024

The left join function in dplyr allows for the merging of two data sets based on a common variable. By specifying selected columns, the left join will only include those specific columns from both data sets in the final merged output. This allows for a more streamlined and efficient way to combine data while still retaining the desired columns. The syntax for performing a left join with selected columns in dplyr is 'left_join(x, y, by = "common variable", select = c("selected columns"))'. This ensures that the output will only contain the specified columns from both data sets, making it easier to analyze and manipulate the merged data.

Perform Left Join Using Selected Columns in dplyr

You can use the following basic syntax in dplyr to perform a left join on two data frames using only selected columns:

```
library(dplyr)

final_df <- df_A %>%
left_join(select(df_B, team, conference), by="team")
```

This particular example will perform a left join on the data frames called df_A and df_B, joining on the column called team, but only the team and conference columns from df_B will be included in the resulting data frame.

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

Example: Perform Left Join Using Selected Columns in dplyr

Suppose we have the following two data frames in R:

```
#create first data frame
df_A <- data.frame(team=c('A', 'B', 'C', 'D', 'E'),
points=c(22, 25, 19, 14, 38))

df_A

team points
1 A 22
2 B 25
3 C 19
4 D 14
5 E 38

#create second data frame
df_B <- data.frame(team=c('A', 'C', 'D', 'F', 'G'),
conference=c('W', 'W', 'E', 'E', 'E'),
rebounds=c(14, 8, 8, 6, 9),
assists=c(4, 3, 9, 9, 4))

df_B

team conference rebounds assists
```

1 A W 14 4

2 C W 8 3

3 D E 8 9

4 F E 6 9

5 G E 9 4

We can use the following syntax in dplyr to perform a left join but only bring in columns team and conference from df_B:

```
library(dplyr)

#perform left join but only bring in team and conference columns from df_B
final_df <- df_A %>%
left_join(select(df_B, team, conference), by="team")

#view final data frame
final_df
```

team points conference

1 A 22 W

2 B 25 NA

3 C 19 W

4 D 14 E

5 E 38 NA

The resulting data frame contains all rows from df_A and only the rows in df_B where the team values matched.

By using the select() function from dplyr, we were able to specify that we only wanted to bring in the team and conference columns from df_B.

Notice that the rebounds and assists columns from df_B were not included in the final data frame.

Note: You can find the complete documentation for the left_join() function in dplyr .