How can I use the anti_join function in dplyr to identify unmatched records between two datasets?

Authored by

stats writer

May 6, 2024

The anti_join function in dplyr is a useful tool for identifying unmatched records between two datasets. This function compares two datasets and returns only the rows that are present in the first dataset but not in the second. It is particularly helpful in identifying missing or inconsistent data between datasets. By using this function, one can easily identify and analyze the unmatched records, which can provide valuable insights for data cleaning and analysis. The anti_join function is a simple yet powerful tool in dplyr that aids in efficiently managing and manipulating data.

dplyr: Use anti_join to Find Unmatched Records

You can use the anti_join() function from the package in R to return all rows in one data frame that do not have matching values in another data frame.

This function uses the following basic syntax:

anti_join(df1, df2, by='col_name')

The following examples show how to use this syntax in practice.

Example 1: Use anti_join() with One Column

Suppose we have the following two data frames in R:

```
#create data frames
df1 <- data.frame(team=c('A', 'B', 'C', 'D', 'E'),
points=c(12, 14, 19, 24, 36))

df2 <- data.frame(team=c('A', 'B', 'C', 'F', 'G'),
```

points=c(12, 14, 19, 33, 17))

We can use the anti_join() function to return all rows in the first data frame that do not have a matching team in the second data frame:

```
library(dplyr)

#perform anti join using 'team' column
anti_join(df1, df2, by='team')
```

```
team points
1 D 24
2 E 36
```

We can see that there are exactly two teams from the first data frame that do not have a matching team name in the second data frame.

Example 2: Use anti_join() with Multiple Columns

Suppose we have the following two data frames in R:

```
#create data frames
df1 <- data.frame(team=c('A', 'A', 'A', 'B', 'B', 'B'),
position=c('G', 'G', 'F', 'G', 'F', 'C'),
```

points=c(12, 14, 19, 24, 36, 41))

df2 <- data.frame(team=c('A', 'A', 'A', 'B', 'B', 'B'),
position=c('G', 'G', 'C', 'G', 'F', 'F'),
points=c(12, 14, 19, 33, 17, 22))

We can use the anti_join() function to return all rows in the first data frame that do not have a matching team *and* position in the second data frame:

library(dplyr)

#perform anti join using 'team' and 'position' columns
anti_join(df1, df2, by=c('team', 'position'))

team position points
1 A F 19
2 B C 41

We can see that there are exactly two records from the first data frame that do not have a matching team name *and* position in the second data frame.