How can we count the distinct values in a dataset using the dplyr package, and could you provide some examples?

Authored by

stats writer

May 12, 2024

The dplyr package is a powerful tool for data manipulation and analysis in R. One common task in data analysis is counting the distinct values in a dataset. This can be easily achieved using the distinct() function in the dplyr package. This function returns a new dataset with only unique combinations of the selected variables. For instance, if we have a dataset with a column of countries, the distinct() function will return a list of all unique countries in that column. This can be useful for identifying unique categories in a dataset or for creating summary tables. Some examples of using the distinct() function include counting the number of unique customers in a sales dataset or the number of unique products in a inventory dataset. Overall, the dplyr package provides a convenient and efficient way to count distinct values in a dataset.

Count Distinct Values Using dplyr (With Examples)

You can use one of the following methods to count the number of distinct values in an R data frame using the n_distinct() function from :

Method 1: Count Distinct Values in One Column

n_distinct(df$column_name)

Method 2: Count Distinct Values in All Columns

sapply(df, function(x) n_distinct(x))

Method 3: Count Distinct Values by Group

df %>% group_by(grouping_column) %>% summarize(count_distinct = n_distinct(values_column))

The following examples show how to use each of these methods in practice with the following data frame:

```
library(dplyr)

#create data frame
df <- data.frame(team=c('A', 'A', 'A', 'A', 'B', 'B', 'B', 'B'),
points=c(6, 6, 8, 10, 9, 9, 12, 12),
assists=c(3, 6, 4, 2, 4, 5, 5, 9))

#view data frame
df

team points assists
1 A 6 3
2 A 6 6
3 A 8 4
4 A 10 2
5 B 9 4
6 B 9 5
7 B 12 5
8 B 12 9
```

Method 1: Count Distinct Values in One Column

The following code shows how to use n_distinct() to

count the number of distinct values in the 'team' column:

#count distinct values in 'team' column
n_distinct(df$team)

2

There are 2 distinct values in the 'team' column.

Method 2: Count Distinct Values in All Columns

The following code shows how to use the sapply() and n_distinct() functions to count the number of distinct values in each column of the data frame:

#count distinct values in every column
sapply(df, function(x) n_distinct(x))
team points assists
2 5 6

From the output we can see:

There are 2 distinct values in the 'team' columnThere are 5 distinct values in the 'points' columnThere are 6 distinct values in the 'assists' column

Method 3: Count Distinct Values by Group

The following code shows how to use the n_distinct() function to count the number of distinct values by group:

```
#count distinct 'points' values by 'team'
df %>%
group_by(team) %>%
summarize(distinct_points = n_distinct(points))

# A tibble: 2 x 2
team distinct_points
1 A 3
2 B 2
```

From the output we can see:

There are 3 distinct points values for team A.There are 2 distinct points values for team B.

The following tutorials explain how to perform other common operations using dplyr: