How can I use the Pandas GroupBy function with the nlargest() method?

Authored by

stats writer

June 29, 2024

The Pandas GroupBy function allows for the grouping of data based on a specific column or set of columns. This function can be further enhanced by using the nlargest() method, which allows for the selection of the top n number of rows based on a specified column or set of columns within each group. This combination of the GroupBy function and nlargest() method provides a powerful tool for analyzing and manipulating data in a structured and organized manner.

Pandas: Use GroupBy with nlargest()

You can use the following syntax to display the n largest values by group in a pandas DataFrame:

```
#display two largest values by group
df.groupby('group_var').nlargest(2)
```

And you can use the following syntax to perform some operation (like taking the sum) on the n largest values by group in a pandas DataFrame:

```
#find sum of two largest values by group
df.groupby('group_var').apply(lambda grp: grp.nlargest(2).sum())
```

The following examples shows how to use each method in practice with the following pandas DataFrame:

```
import pandas as pd
```

#create DataFrame
df = pd.DataFrame({'team': ,
'points': })

#view DataFrame
print(df)

team points
0 A 12
1 A 29
2 A 34
3 A 14
4 A 10
5 B 11
6 B 7
7 B 36
8 B 34
9 B 22

Example 1: Display N Largest Values by Group

We can use the following syntax to display the two largest points values grouped by team:

#display two largest points values grouped by team

```
df.groupby('team').nlargest(2)

team
A 2 34
1 29
B 7 36
8 34
Name: points, dtype: int64
```

The output shows the two largest points values for each team, along with their index positions in the original DataFrame.

Example 2: Perform Operation on N Largest Values by Group

We can use the following syntax to calculate the sum of the two largest points values grouped by team:

```
#calculate sum of two largest points values for each team
df.groupby('team').apply(lambda grp: grp.nlargest(2).sum())

team
A 63
B 70
```

Name: points, dtype: int64

Here's how to interpret the output:

The sum of the two largest points values for team A is 63.The sum of the two largest points values for team B is 70.

We can use similar syntax to calculate the mean of the two largest points values grouped by team:

```
#calculate mean of two largest points values for each team
df.groupby('team').apply(lambda grp: grp.nlargest(2).mean())

team
A 31.5
B 35.0
Name: points, dtype: float64
```

The mean of the two largest points values for team A is 31.5.The mean of the two largest points values for team B is 35.0.

Note: You can find the complete documentation for the GroupBy function .

Additional Resources

The following tutorials explain how to perform other common operations in pandas: