How can I use groupby with diff- in Pandas?

Authored by

stats writer

June 26, 2024

Groupby with diff- in Pandas is a function that allows for the grouping of data in a DataFrame based on a specific criteria, and then performing a specified calculation on the grouped data. This function is useful for analyzing and manipulating large datasets, as it allows for the efficient comparison and calculation of data within the groups. By using the diff- function, data can be grouped by a particular column or variable, and then the difference between values within each group can be calculated. This can provide valuable insights and aid in data analysis and decision making.

Pandas: Use groupby with diff

You can use the following basic syntax to use the groupby() function with the diff() function in pandas:

df = df.sort_values(by=)

df = df.groupby().diff().fillna(0)

This particular example sorts the rows of the DataFrame by two specific variables, then groups by group_var1 and calculates the difference between rows in the values_var column.

Note that fillna(0) tells pandas to insert a zero whenever the value of the group variable changes between consecutive rows in the DataFrame.

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

Example: How to Use groupby with diff in Pandas

Suppose we have the following pandas DataFrame that contains the total sales made by two different stores on various dates:

```
import pandas as pd

#create DataFrame
df = pd.DataFrame({'store': ,
'date': pd.to_datetime(),
'sales': })

#view DataFrame
print(df)
```

```
store date sales
0 A 2022-01-01 12
1 A 2022-01-02 15
2 A 2022-01-03 24
3 A 2022-01-04 24
4 B 2022-01-01 14
5 B 2022-01-02 19
6 B 2022-01-03 12
7 B 2022-01-04 38
```

Now suppose that we would like to create a new column called sales_diff that contains the difference in sales values between consecutive dates, grouped by store.

We can use the following syntax to do so:

```
#sort DataFrame by store and date
df = df.sort_values(by=)

#create new column that contains difference between sales grouped by store
df = df.groupby().diff().fillna(0)

#view update DataFrame
print(df)
```

```
store date sales sales_diff
0 A 2022-01-01 12 0.0
1 A 2022-01-02 15 3.0
2 A 2022-01-03 24 9.0
3 A 2022-01-04 24 0.0
4 B 2022-01-01 14 0.0
5 B 2022-01-02 19 5.0
6 B 2022-01-03 12 -7.0
7 B 2022-01-04 38 26.0
```

The new sales_diff column contains the difference in sales values between consecutive dates, grouped by store.

For example, we can see:

The difference in sales at store A between 1/1/2022 and 1/2/2022 is 3.The difference in sales at store A between 1/2/2022 and 1/3/2022 is 9.The difference in sales at store A between 1/3/2022 and 1/4/2022 is 0.

And so on.

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