How can I group a Pandas DataFrame by year, and provide an example?

Authored by

stats writer

June 26, 2024

Grouping a Pandas DataFrame by year refers to the process of organizing data from a DataFrame based on the year it corresponds to. This can be achieved by using the "groupby" function in Pandas, which allows for grouping data by a specific column or attribute. For example, if we have a DataFrame containing sales data for a company from multiple years, we can use the "groupby" function to group the data by year and calculate the total sales for each year. This allows for easy analysis and comparison of data between different years.

Group by Year in Pandas DataFrame (With Example)

You can use the following basic syntax to group rows by year in a pandas DataFrame:

df.groupby(df.your_date_column.dt.year).sum()

This particular formula groups the rows by date in your_date_column and calculates the sum of values for the values_column in the DataFrame.

Note that the dt.year() function extracts the year from a date column in pandas.

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

Example: How to Group by Year in Pandas

Suppose we have the following pandas DataFrame that shows the sales made by some company on various

dates:

```
import pandas as pd

#create DataFrame
df = pd.DataFrame({'date':
pd.date_range(start='1/1/2020', freq='3m', periods=10),
'sales': ,
'returns': })

#view DataFrame
print(df)
```

date sales returns
0 2020-01-31 6 0
1 2020-04-30 8 3
2 2020-07-31 9 2
3 2020-10-31 11 2
4 2021-01-31 13 1
5 2021-04-30 8 3
6 2021-07-31 8 2
7 2021-10-31 15 4
8 2022-01-31 22 1
9 2022-04-30 9 5

Related:

We can use the following syntax to calculate the sum of sales grouped by year:

```
#calculate sum of sales grouped by year
df.groupby(df.date.dt.year).sum()
```

```
date
2020 34
2021 44
2022 31
Name: sales, dtype: int64
```

Here's how to interpret the output:

The total sales made during 2020 was 34.The total sales made during 2021 was 44.The total sales made during 2022 was 31.

We can use similar syntax to calculate the max of the sales values grouped by year:

```
#calculate max of sales grouped by year
df.groupby(df.date.dt.year).max()
```

date

2020 11

2021 15

2022 22

Name: sales, dtype: int64

We can use similar syntax to calculate any value we'd like grouped by the year value of a date column.

Note: You can find the complete documentation for the GroupBy operation in pandas .

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