How can I drop duplicates in a Pandas dataframe while keeping only the latest entries?

Authored by

stats writer

June 25, 2024

This process involves using the Pandas library in Python to drop duplicate rows in a dataframe while retaining only the most recent entries. By using the "drop_duplicates" function and specifying the "keep" parameter as "last", the dataframe will be modified to remove any duplicate rows and only keep the most recent entries based on a specified column. This allows for a more streamlined and accurate analysis of data without the interference of duplicate entries.

Pandas: Drop Duplicates and Keep Latest

You can use the following basic syntax to drop duplicates from a pandas DataFrame but keep the row with the latest timestamp:

df = df.sort_values('time').drop_duplicates(, keep='last')

This particular example drops rows with duplicate values in the item column, but keeps the row with the latest timestamp in the time column.

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

Example: Drop Duplicates and Keep Latest in Pandas

Suppose we have the following pandas DataFrame that contains information about the sales of various fruits at some grocery store:

import pandas as pd

```
#create DataFrame
df = pd.DataFrame({'time': ,
'item': ,
'sales': })

#convert time column to datetime dtype
df = pd.to_datetime(df)

#view DataFrame
print(df)

time item sales
0 2022-10-25 04:00:00 apple 18
1 2022-10-25 11:55:12 orange 22
2 2022-10-26 02:00:00 apple 19
3 2022-10-27 10:30:00 mango 14
4 2022-10-27 14:25:00 mango 14
5 2022-10-28 01:15:27 kiwi 11
```


Suppose we would like to remove all rows with duplicate values in the item column but keep the row with the latest timestamp in the time column.

We can use the following syntax to do so:

#drop duplicate rows based on value in 'item' column but keep latest timestamp
df = df.sort_values('time').drop_duplicates(, keep='last')

#view updated DataFrame
print(df)

time item sales
1 2022-10-25 11:55:12 orange 22
2 2022-10-26 02:00:00 apple 19
4 2022-10-27 14:25:00 mango 14
5 2022-10-28 01:15:27 kiwi 11

Notice that the item column had multiple rows with 'apple' and 'mango' as values.

Each of these duplicate rows were removed but the row with the latest timestamp in the time column was kept.

Note: If you would like to remove rows based on duplicate values in multiple columns, simply include multiple column names in the first argument of the drop_duplicates() function.

The following tutorials explain how to perform other

common tasks in pandas: