How can I find duplicates in a Pandas DataFrame and remove them?

Authored by

stats writer

July 1, 2024

To find and remove duplicates in a Pandas DataFrame, one can use the "drop_duplicates" method. This method allows for the identification and removal of rows in a DataFrame that contain identical values. By specifying the columns to be checked for duplicates, the method can accurately identify and remove any duplicate rows. This process can be repeated for multiple columns or the entire DataFrame, ensuring that all duplicates are removed. This method is a simple and efficient way to clean up data and ensure accuracy in analysis.

Find Duplicates in Pandas DataFrame (With Examples)

You can use the function to find duplicate values in a pandas DataFrame.

This function uses the following basic syntax:

#find duplicate rows across all columns
duplicateRows = df

#find duplicate rows across specific columns
duplicateRows = df)]

The following examples show how to use this function in practice with the following pandas DataFrame:

import pandas as pd

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

'assists': })

#view DataFrame

print(df)

team points assists

0 A 10 5

1 A 10 5

2 A 12 7

3 A 12 9

4 B 15 12

5 B 17 9

6 B 20 6

7 B 20 6

Example 1: Find Duplicate Rows Across All Columns

The following code shows how to find duplicate rows across all of the columns of the DataFrame:

#identify duplicate rows

duplicateRows = df

#view duplicate rows

duplicateRows

team points assists
1 A 10 5
7 B 20 6

There are two rows that are exact duplicates of other rows in the DataFrame.

Note that we can also use the argument keep='last' to display the first duplicate rows instead of the last:

#identify duplicate rows
duplicateRows = df

#view duplicate rows
print(duplicateRows)

team points assists
0 A 10 5
6 B 20 6

Example 2: Find Duplicate Rows Across Specific Columns

The following code shows how to find duplicate rows across just the 'team' and 'points' columns of the DataFrame:

#identify duplicate rows across 'team' and 'points' columns
duplicateRows = df)]

#view duplicate rows
print(duplicateRows)

team points assists
1 A 10 5
3 A 12 9
7 B 20 6

There are three rows where the values for the 'team' and 'points' columns are exact duplicates of previous rows.

Example 3: Find Duplicate Rows in One Column

The following code shows how to find duplicate rows in just the 'team' column of the DataFrame:

#identify duplicate rows in 'team' column
duplicateRows = df)]

#view duplicate rows
print(duplicateRows)

team points assists
1 A 10 5
2 A 12 7
3 A 12 9
5 B 17 9
6 B 20 6
7 B 20 6

There are six total rows where the values in the 'team' column are exact duplicates of previous rows.

Additional Resources

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