

How can I remove duplicate rows from a Pandas DataFrame?

Authored by
stats writer

April 23, 2024

RECOMMENDED CITATION

stats writer (2024). *How can I remove duplicate rows from a Pandas DataFrame?*.
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=138505>

Removing duplicate rows from a Pandas DataFrame involves identifying and deleting rows that have identical values in all columns. This process can be achieved by using the `drop_duplicates()` function, which allows for various parameters to be specified such as the columns to consider and the method for determining duplicates. By utilizing this function, duplicate rows can be efficiently removed from a DataFrame, ensuring data integrity and accuracy.

Drop Duplicate Rows in a Pandas DataFrame

The easiest way to drop duplicate rows in a pandas DataFrame is by using the function, which uses the following syntax:

```
df.drop_duplicates(subset=None, keep='first',  
inplace=False)
```

where:

subset: Which columns to consider for identifying duplicates. Default is all columns.
keep: Indicates which duplicates (if any) to keep. **first:** Delete all duplicate rows except first. **last:** Delete all duplicate rows except last. **False:** Delete all duplicates.
inplace: Indicates whether to drop duplicates in place or return a copy of the DataFrame.

This tutorial provides several examples of how to use this function in practice on the following DataFrame:

```
import pandas as pd

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

```
#display DataFrame
print(df)
```

```
team points assists
0 a 3 8
1 b 7 6
2 b 7 7
3 c 8 9
4 c 8 9
5 d 9 3
```

Example 1: Remove Duplicates Across All Columns

The following code shows how to remove rows that have duplicate values across *all* columns:

```
df.drop_duplicates()

team points assists
```

0 a 3 8

1 b 7 6

2 b 7 7

3 c 8 9

5 d 9 3

By default, the `drop_duplicates()` function deletes all duplicates except the first.

However, we could use the `keep=False` argument to delete all duplicates entirely:

```
df.drop_duplicates(keep=False)
```

team points assists

0 a 3 8

1 b 7 6

2 b 7 7

5 d 9 3

Example 2: Remove Duplicates Across Specific Columns

The following code shows how to remove rows that have duplicate values across just the columns titled *team* and *points*:

df.drop_duplicates(subset=)

team points assists

0 a 3 8

1 b 7 6

3 c 8 9

5 d 9 3

How to Drop Duplicate Columns in Pandas