how do I remove rows with duplicate values of columns in pandas data frame?

Ask Question

3

I have a pandas data frame which looks like this.

```
'Column1' 'Column2' 'Column3'
'cat'     'bat'.    'xyz'
'toy'    'flower'.  'abc'
'cat'    'bat'      'lmn'
```

I want to identify that cat and bat are same values which have been repeated and hence want to remove one record and preserve only the first record. The resulting data frame should only have.

```
'Column1'  'Column2' 'Column3'
'cat'.     'bat'.     'xyz'
'toy'.     'flower'.  'abc'
```

python pandas

edited Jun 16 '18 at 5:07

asked Jun 16 '18 at 4:57

Sayonti
25 2 9

1 `df.drop_duplicates(['Column1', 'Column2'])` – piRSquared Jun 16 '18 at 5:01

I am looking for something that will match the values in the two particular columns and then drop not for the entire data frame @piRSquared – Sayonti Jun 16 '18 at 5:04

Did you look into `subset` option in `drop_duplicates` ? – student Jun 16 '18 at 5:08

2 you need `keep='first'` which is the default. `keep=False` is wrong – piRSquared Jun 16 '18 at 5:14

docs/stable/generated/… Also, you are using duplicated which only keeps duplcates, instead need drop_duplicates. pandas.pydata.org/pandas-docs/stable/generated/… – student Jun 16 '18 at 5:15

3 Answers

5

✓

Using `drop_duplicates` with `subset` with list of columns to check for duplicates on and `keep='first'` to keep first of duplicates.

If `dataframe` is:

```
df = pd.DataFrame({'Column1': ["'
                   'Column2': ["'
                   'Column3': ["'
print(df)
```

Result:

```
  Column1   Column2 Column3
0   'cat'     'bat'   'xyz'
1   'toy'  'flower'   'abc'
2   'cat'     'bat'   'lmn'
```

Then:

```
result_df = df.drop_duplicates(su
print(result_df)
```

Result:

```
  Column1   Column2 Column3
0   'cat'     'bat'   'xyz'
1   'toy'  'flower'   'abc'
```

answered Jun 16 '18 at 5:29

student

8,515 3 16 31

0

```
import pandas as pd

df = pd.DataFrame({"Column1":["cat
                   "Column2":[1,1
                   "Column3":["C"

df = df.drop_duplicates(subset=['C
print(df)
```

111 1 7

add 'Column2' as well inside subset parameter. – Jay Dangar Jun 16 '18 at 5:46

While this code snippet may be the solution, including an explanation really helps to improve the quality of your post. Remember that you are answering the question for readers in the future, and those people might not know the reasons for your code suggestion. – Narendra Jadhav Jun 16 '18 at 6:47

I agree. I will try to do that. Thanks, Narendra! – zafrin Jun 16 '18 at 14:41

0

Inside the `drop_duplicates()` method of `Dataframe` you can provide a series of column names to eliminate duplicate records from your data.

The following "Tested" code does the same :

```
import pandas as pd

df = pd.DataFrame()
df.insert(loc=0,column='Column1',v
df.insert(loc=1,column='Column2',v
df.insert(loc=2,column='Column3',v

df = df.drop_duplicates(subset=['C
print(df)
```

Inside of the subset parameter, you can insert other column names as well and by default it will consider all the columns of your data and you can provide keep value as :-

- first : Drop duplicates except for the first occurrence.
- last : Drop duplicates except for the last occurrence.
- False : Drop all duplicates.

edited Jun 16 '18 at 9:14

Mr. T
4,223 9 16 36

answered Jun 16 '18 at 5:35

Jay Dangar
755 5 18