How can I sort a Pandas DataFrame based on a string column?

Authored by

stats writer

June 25, 2024

RECOMMENDED CITATION

stats writer (2024). *How can I sort a Pandas DataFrame based on a string column?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=152203

ARABPSYCHOLOGY.COM

Sorting a Pandas DataFrame based on a string column can be done using the "sort_values()" function. This function allows the user to specify the column to be sorted and the desired sorting order. The DataFrame will be sorted in ascending order by default, but the "ascending" parameter can be set to False to sort in descending order. It is important to note that the "inplace" parameter should be set to True if the changes are to be made to the original DataFrame, otherwise a new sorted DataFrame will be returned. By using the "sort_values()" function, it is possible to easily organize and arrange a DataFrame based on the values in a specific string column.

Pandas: Sort DataFrame Based on String Column

You can use the following methods to sort the rows of a pandas DataFrame based on the values in a particular string column:

Method 1: Sort by String Column (when column only contains characters)

```
df = df.sort_values('my_string_column')
```

Method 2: Sort by String Column (when column contains characters *and* digits)

```
#create 'sort' column that contains digits from 'my_string_column'
df = df.str.extract('(d+)', expand=False).astype(int)
#sort rows based on digits in 'sort' column
df = df.sort_values('sort')
```

ARABPSYCHOLOGY.COM

The following examples show how to use each method in practice.

Example 1: Sort by String Column (when column only contains characters)

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

```
import pandas as pd

#create DataFrame
df = pd.DataFrame({'product': ,
'sales': })

#view DataFrame
print(df)
```

product sales
0 Apples 18
1 Oranges 22
2 Bananas 19
3 Lettuce 14
4 Beans 29

ARABPSYCHOLOGY.COM

We can use the following syntax to sort the rows of the DataFrame based on the strings in the product column:

```
#sort rows from A to Z based on string in 'product' column
df = df.sort_values('product')

#view updated DataFrame
print(df)
```

```
  product  sales
0  Apples     18
2 Bananas     19
4   Beans     29
3 Lettuce     14
1 Oranges     22
```

Notice that the rows are now sorted from A to Z based on the strings in the product column.

If you'd like to instead sort from Z to A, simply add the argument ascending=False:

```
#sort rows from Z to A based on string in 'product' column
```

ARABPSYCHOLOGY.COM

df = df.sort_values('product', ascending=False)

#view updated DataFrame
print(df)

product sales
1 Oranges 22
3 Lettuce 14
4 Beans 29
2 Bananas 19
0 Apples 18

Notice that the rows are now sorted from Z to A based on the strings in the product column.

Example 2: Sort by String Column (when column contains characters *and* digits)

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

import pandas as pd

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

ARABPSYCHOLOGY.COM

```
'sales': })

#view DataFrame
print(df)

product sales
0 A3 18
1 A5 22
2 A22 19
3 A50 14
4 A2 14
5 A7 11
6 A9 20
7 A13 28
```

Notice that the strings in the product column contain both characters and digits.

If we attempt to sort the rows of the DataFrame using the values in the product column, the strings will not be sorted in the correct order based on the digits:

```
import pandas as pd

#sort rows based on strings in 'product' column
```

ARABPSYCHOLOGY.COM

```
df = df.sort_values('product')

#view updated DataFrame
print(df)

product sales
7 A13 28
4 A2 14
2 A22 19
0 A3 18
1 A5 22
3 A50 14
5 A7 11
6 A9 20
```

Instead, we must create a new temporary column called sort that contains only the digits from the product column, then sort by the values in the sort column, then drop the column entirely:

```
import pandas as pd

#create new 'sort' column that contains digits from 'product' column
df = df.str.extract('(d+)', expand=False).astype(int)
```

ARABPSYCHOLOGY.COM

```
#sort rows based on digits in 'sort' column
df = df.sort_values('sort')

#drop 'sort' column
df = df.drop('sort', axis=1)

#view updated DataFrame
print(df)

product sales
4 A2 14
0 A3 18
1 A5 22
5 A7 11
6 A9 20
7 A13 28
2 A22 19
3 A50 14
```

ARABPSYCHOLOGY.COM

Notice that the rows are now sorted by the strings in the product column and the digits are sorted in the correct order.

Note: You can find the complete documentation for the pandas sort_values() function .

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

ARABPSYCHOLOGY.COM