How can I extract a substring from an entire column in a Pandas dataframe?

Authored by

stats writer

June 26, 2024

Extracting a substring from an entire column in a Pandas dataframe refers to the process of retrieving a specific section of text or characters from a column in a dataframe. This can be achieved by using the built-in functions and methods in Pandas, such as the .str.extract() method. This allows for the extraction of data based on a certain pattern or condition, making it a useful tool for data manipulation and analysis. By specifying the desired substring and the column to extract from, users can easily retrieve the necessary information from their dataframe.

Pandas: Get Substring of Entire Column

You can use the following basic syntax to get the substring of an entire column in a pandas DataFrame:

df = df.str

This particular example creates a new column called some_substring that contains the characters from positions 1 through 4 in the string_column.

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

Example: Get Substring of Entire Column in Pandas

Suppose we have the following pandas DataFrame that contains information about various basketball teams:

import pandas as pd

#create DataFrame

df = pd.DataFrame({'team': ,
'points': })

#view DataFrame
print(df)

team points
0 Mavericks 120
1 Warriors 132
2 Rockets 108
3 Hornets 118
4 Lakers 106

We can use the following syntax to create a new column that contains the characters in the team column between positions 1 and 4:

#create column that extracts characters in positions 1 through 4 in team column
df = df.str

#view updated DataFrame
print(df)

team points team_substring

0 Mavericks 120 ave
1 Warriors 132 arr
2 Rockets 108 ock
3 Hornets 118 orn
4 Lakers 106 ake

The new column called team_substring contains the characters in the team column between positions 1 and 4.

Note that if you attempt to use this syntax to extract a substring from a numeric column, you'll receive an error:

#attempt to extract characters in positions 0 through 2 in points column
df = df.str

AttributeError: Can only use .str accessor with string values!

Instead, you must convert the numeric column to a string by using astype(str) first:

#extract characters in positions 0 through 2 in points

column
df = df.astype(str).str

#view updated DataFrameprint(df)

team points points_substring
0 Mavericks 120 12
1 Warriors 132 13
2 Rockets 108 10
3 Hornets 118 11
4 Lakers 106 10

This time we're able to successfully extract characters in positions 0 through 2 of the points column because we first converted it to a string.

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