How can I use Pandas to sample rows with replacement in a dataset?

Authored by

stats writer

June 24, 2024

Pandas is a powerful Python library that is commonly used for data analysis and manipulation. One of its useful functions is the ability to sample rows with replacement in a dataset. This means that when selecting a random subset of rows from a dataset, the same row can be selected multiple times. This can be done by using the "sample" function in Pandas, which allows for specifying the number of rows to be sampled as well as the option for replacement. By using this function, analysts can easily obtain multiple random samples from a dataset, which can be useful for statistical analysis and modeling.

Pandas: Sample Rows with Replacement

You can use the argument replace=True within the pandas sample() function to randomly sample rows in a DataFrame with replacement:

#randomly select *n* rows with repeats allowed
df.sample(n=5, replace=True)

By using replace=True, you allow the same row to be included in the sample multiple times.

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

Example: Sample Rows with Replacement in Pandas

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

import pandas as pd

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

#view DataFrame
print(df)

team points assists rebounds
0 A 18 5 11
1 B 22 7 8
2 C 19 7 10
3 D 14 9 6
4 E 14 12 6
5 F 11 9 5
6 G 20 9 9
7 H 28 4 12

Suppose we use the sample() function to randomly select a sample of rows:

#randomly select 6 rows from DataFrame (without replacement)

```
df.sample(n=6, random_state=0)

team points assists rebounds
6 G 20 9 9
2 C 19 7 10
1 B 22 7 8
7 H 28 4 12
3 D 14 9 6
0 A 18 5 11
```

Notice that six rows have been selected from the DataFrame and none of the rows appear multiple times in the sample.

Note: The argument random_state=0 ensures that this example is reproducible.

Now suppose we use the argument replace=True to select a random sample of rows with replacement:

```
#randomly select 6 rows from DataFrame (with replacement)
df.sample(n=6, replace=True, random_state=0)

team points assists rebounds
```

4 E 14 12 6

7 H 28 4 12

5 F 11 9 5

0 A 18 5 11

3 D 14 9 6

3 D 14 9 6

Notice that the row with team "D" appears multiple times.

By using the argument replace=True, we allow the same row to appear in the sample multiple times.

Also note that we could select a random fraction of the DataFrame to be included in the sample by using the frac argument.

For example, the following example shows how to select 75% of rows to be included in the sample with replacement:

#randomly select 75% of rows (with replacement)
df.sample(frac=0.75, replace=True, random_state=0)

team points assists rebounds

4 E 14 12 6

7 H 28 4 12

5 F 11 9 5

0 A 18 5 11

3 D 14 9 6

3 D 14 9 6

Notice that 75% of the number of rows (6 out of 8) were included in the sample and at least one of the rows (with team "D") appeared in the sample twice.

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

The following tutorials explain how to perform other common sampling methods in Pandas: