How can I calculate summary statistics for a Pandas DataFrame?

Authored by

stats writer

June 30, 2024

Pandas is a popular library in Python used for data analysis and manipulation. It provides a variety of tools for calculating summary statistics on data stored in a DataFrame, which is a tabular data structure. To calculate summary statistics for a Pandas DataFrame, one can use the built-in functions such as .describe() and .mean() which provide information on measures like mean, median, standard deviation, and more. Additionally, users can also use custom functions and methods to calculate specific summary statistics based on their data needs. With its versatile features, Pandas offers a convenient and efficient way to calculate summary statistics for a DataFrame, making it a valuable tool for data analysis.

Calculate Summary Statistics for a Pandas DataFrame

You can use the following methods to calculate summary statistics for variables in a pandas DataFrame:

Method 1: Calculate Summary Statistics for All Numeric Variables

df.describe()

Method 2: Calculate Summary Statistics for All String Variables

df.describe(include='object')

Method 3: Calculate Summary Statistics Grouped by a Variable

df.groupby('group_column').mean()

df.groupby('group_column').median()

df.groupby('group_column').max()

...

The following examples show how to use each method in practice with the following pandas DataFrame:

```
import pandas as pd
import numpy as np

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

#view DataFrame
print(df)

team points assists rebounds
0 A 18 5.0 11.0
1 A 22 NaN 8.0
```

2 A 19 7.0 10.0

3 A 14 9.0 6.0

4 B 14 12.0 6.0

5 B 11 9.0 5.0

6 B 20 9.0 9.0

7 B 28 4.0 NaN

8 B 30 5.0 6.0

Example 1: Calculate Summary Statistics for All Numeric Variables

The following code shows how to calculate the summary statistics for each numeric variable in the DataFrame:

df.describe()

points assists rebounds

count 9.000000 8.000000 8.000000

mean 19.555556 7.500000 7.625000

std 6.366143 2.725541 2.199838

min 11.000000 4.000000 5.000000

25% 14.000000 5.000000 6.000000

50% 19.000000 8.000000 7.000000

75% 22.000000 9.000000 9.250000

max 30.000000 12.000000 11.000000

We can see the following summary statistics for each of the three numeric variables:

count: The count of non-null valuesmean: The mean valuestd: The standard deviationmin: The minimum value25%: The value at the 25th percentile50%: The value at the 50th percentile (also the median)75%: The value at the 75th percentilemax: The maximum value

Example 2: Calculate Summary Statistics for All String Variables

The following code shows how to calculate the summary statistics for each string variable in the DataFrame:

df.describe(include='object')

team
count 9
unique 2
top B
freq 5

count: The count of non-null valuesunique: The number of unique valuestop: The most frequently occurring valuefreq: The count of the most frequently occurring

value

Example 3: Calculate Summary Statistics Grouped by a Variable

The following code shows how to calculate the mean value for all numeric variables, grouped by the team variable:

```
df.groupby('team').mean()
```

```
points assists rebounds
team
A 18.25 7.0 8.75
B 20.60 7.8 6.50
```

The output displays the mean value for the points, assists, and rebounds variables, grouped by the team variable.

Note that we can use similar syntax to calculate a different summary statistic, such as the median:

```
df.groupby('team').median()
```

```
points assists rebounds
team
```

A 18.5 7.0 9.0

B 20.0 9.0 6.0

The output displays the median value for the points, assists, and rebounds variables, grouped by the team variable.

Note: You can find the complete documentation for the describe function in pandas .

Additional Resources

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