How can I calculate a five number summary in Pandas?

Authored by

stats writer

June 26, 2024

Calculating a five number summary in Pandas involves using the describe() function, which provides statistical information on a dataset such as minimum, maximum, median, and quartile values. Specifically, the five number summary includes the minimum, first quartile, median, third quartile, and maximum values of a dataset. By using the describe() function, Pandas allows for a quick and efficient way to obtain these summary statistics, making it a useful tool for data analysis and decision making.

Calculate a Five Number Summary in Pandas

A five number summary is a way to summarize a dataset using the following five values:

The minimumThe first quartileThe medianThe third quartileThe maximum

The five number summary is useful because it provides a concise summary of the distribution of the data in the following ways:

It tells us where the middle value is located, using the median.It tells us how spread out the data is, using the first and third quartiles.It tells us the range of the data, using the minimum and the maximum.

The easiest way to calculate a five number summary for variables in a pandas DataFrame is to use the describe() function as follows:

df.describe().loc]

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

Example: Calculate Five Number Summary in Pandas DataFrame

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

We can use the following syntax to calculate the five number summary for each numeric variable in the DataFrame:

#calculate five number summary for each numeric variable
df.describe().loc]

points assists rebounds
min 11.0 4.0 5.00
25% 14.0 6.5 6.00
50% 18.5 8.0 8.50
75% 20.5 9.0 10.25
max 28.0 12.0 12.00

Here's how to interpret the output for the points variable:

The minimum value is 11.The value at the 25th percentile is 14.The value at the 50th percentile is

18.5.The value at the 75th percentile is 20.5.The maximum value is 28.

We can interpret the values for the assists and rebounds variables in a similar manner.

If you'd only like to calculate the five number summary for one specific variable in the DataFrame, you can use the following syntax:

#calculate five number summary for the points variable
df.describe().loc]

min 11.0
25% 14.0
50% 18.5
75% 20.5
max 28.0
Name: points, dtype: float64

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