How does the summary statistics of different groups in a dataset compare when using the describe() function in Pandas?”

Authored by

stats writer

June 27, 2024

The describe() function in Pandas is a powerful tool for analyzing data by providing summary statistics for a given dataset. When comparing the summary statistics of different groups within a dataset, the describe() function allows for a quick and concise comparison. It calculates key metrics such as mean, standard deviation, minimum and maximum values, and quartile values for each group, making it easy to identify any differences or similarities between groups. This allows for a more comprehensive understanding of the data and can assist in making informed decisions based on the data analysis. Overall, the describe() function in Pandas provides a clear and efficient way to compare summary statistics of different groups within a dataset.

Pandas: Use describe() by Group

You can use the describe() function to generate descriptive statistics for variables in a pandas DataFrame.

You can use the following basic syntax to use the describe() function with the groupby() function in pandas:

df.groupby('group_var').describe()

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

Example: Use describe() by Group in Pandas

Suppose we have the following pandas DataFrame that contains information about basketball players on two different teams:

```
import pandas as pd

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

#view DataFrame
print(df)

team points assists
0 A 8 2
1 A 12 2
2 A 14 3
3 A 14 5
4 B 15 7
5 B 22 6
6 B 27 8
7 B 24 12
```


We can use the describe() function along with the groupby() function to summarize the values in the points column for each team:

```
#summarize points by team
```

```
df.groupby('team').describe()
count mean std min 25% 50% 75% max
team
A 4.0 12.0 2.828427 8.0 11.00 13.0 14.00 14.0
B 4.0 22.0 5.099020 15.0 20.25 23.0 24.75 27.0
```

From the output, we can see the following values for the points variable for each team:

count (number of observations)mean (mean points value)std (standard deviation of points values)min (minimum points value)25% (25th percentile of points)50% (50th percentile (i.e. median) of points)75% (75th percentile of points)max (maximum points value)

If you'd like the results to be displayed in a DataFrame format, you can use the reset_index() argument:

```
#summarize points by team
df.groupby('team').describe().reset_index()

team count mean std min 25% 50% 75% max
0 A 4.0 12.0 2.828427 8.0 11.00 13.0 14.00 14.0
1 B 4.0 22.0 5.099020 15.0 20.25 23.0 24.75 27.0
```

The variable team is now a column in the DataFrame and the index values are 0 and 1.

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