How can I calculate correlation by group in Pandas?

Authored by

stats writer

July 1, 2024

Calculating correlation by group in Pandas refers to the process of determining the strength and direction of the relationship between two or more variables within a specific group of data using the Pandas library in Python. This can be achieved by grouping the data based on a categorical variable and then calculating the correlation coefficient between the variables for each group. The resulting correlation values can provide insights into any potential patterns or trends within the data and can help in making informed decisions for further analysis.

Calculate Correlation By Group in Pandas

You can use the following basic syntax to calculate the correlation between two variables by group in pandas:

df.groupby('group_var')].corr().unstack().iloc

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

Example: Calculate Correlation By Group in Pandas

Suppose we have the following pandas DataFrame:

import pandas as pd

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

#view DataFrame

print(df)

We can use the following code to calculate the correlation between points and assists, grouped by team:

#calculate correlation between points and assists, grouped by team
df.groupby('team')].corr().unstack().iloc

team
A 0.603053
B 0.981798
Name: (points, assists), dtype: float64

From the output we can see:

The correlation coefficient between points and assists for team A is .603053.The correlation coefficient between points and assists for team B is .981798.

Since both correlation coefficients are positive, this tells us that the relationship between points and assists for both teams is positive.

That is, players who tend to score more points also tend to record more assists.

Related:

Note that we could shorten the syntax by not using the unstack and iloc functions, but the results are uglier:

df.groupby('team')].corr()

points assists
team
A points 1.000000 0.603053
assists 0.603053 1.000000
B points 1.000000 0.981798
assists 0.981798 1.000000

This syntax produces a correlation matrix for both teams, which provides us with excessive information.

Additional Resources