How can I plot the distribution of column values in Pandas?

Authored by

stats writer

June 27, 2024

To plot the distribution of column values in Pandas, one can use the built-in function "hist()" or "plot(kind='hist')" to generate a histogram. This will display the frequency of values in the chosen column, allowing for a visual representation of the data's distribution. Other options include using the "boxplot()" or "plot(kind='box')" function to create a box plot, which can show the median, quartiles, and outliers of the column's values. One can also utilize the "scatter()" or "plot(kind='scatter')" function to create a scatter plot, which can show the relationship between two columns and the distribution of their values. These various methods in Pandas provide a convenient way to analyze and visualize the distribution of column values in a dataset.

Plot Distribution of Column Values in Pandas

You can use the following methods to plot a distribution of column values in a pandas DataFrame:

Method 1: Plot Distribution of Values in One Column

df.plot(kind='kde')

Method 2: Plot Distribution of Values in One Column, Grouped by Another Column

df.groupby('group_column').plot(kind='kde')

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

import pandas as pd

#create DataFrame

```
df = pd.DataFrame({'team': ,
'points': })

#view DataFrame
print(df)

team points
0 A 3
1 A 3
2 A 4
3 A 5
4 A 4
5 A 7
6 A 7
7 A 7
8 A 10
9 A 11
10 B 8
11 B 7
12 B 8
13 B 9
14 B 12
15 B 12
16 B 12
```

17 B 14

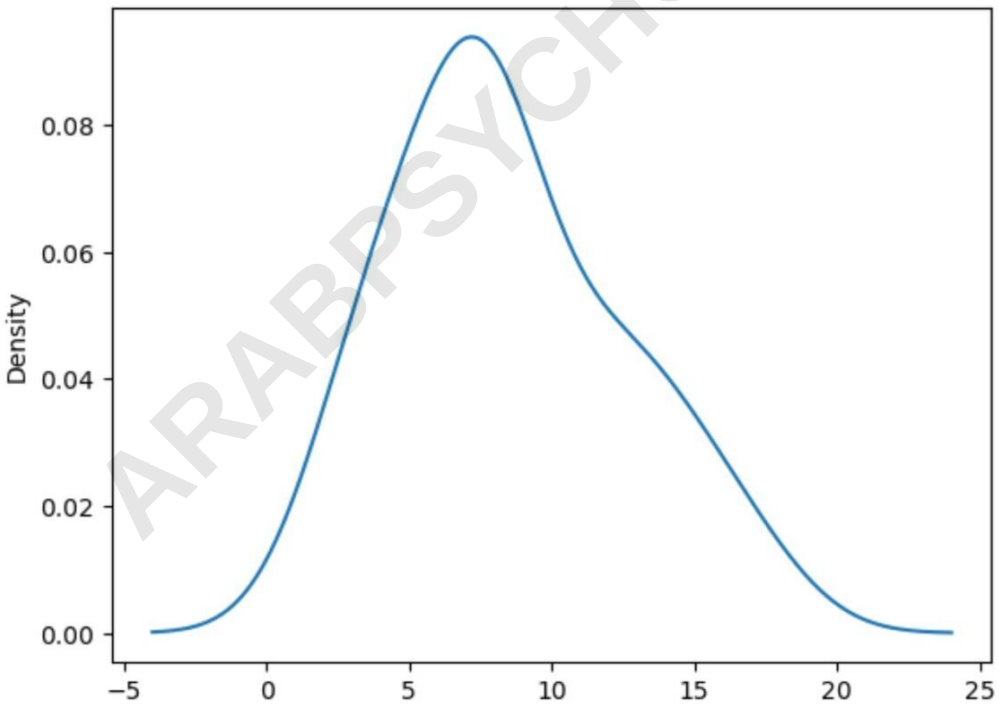
18 B 15

19 B 17

Example 1: Plot Distribution of Values in One Column

The following code shows how to plot the distribution of values in the points column:

```
#plot distribution of values in points column
df.plot(kind='kde')
```

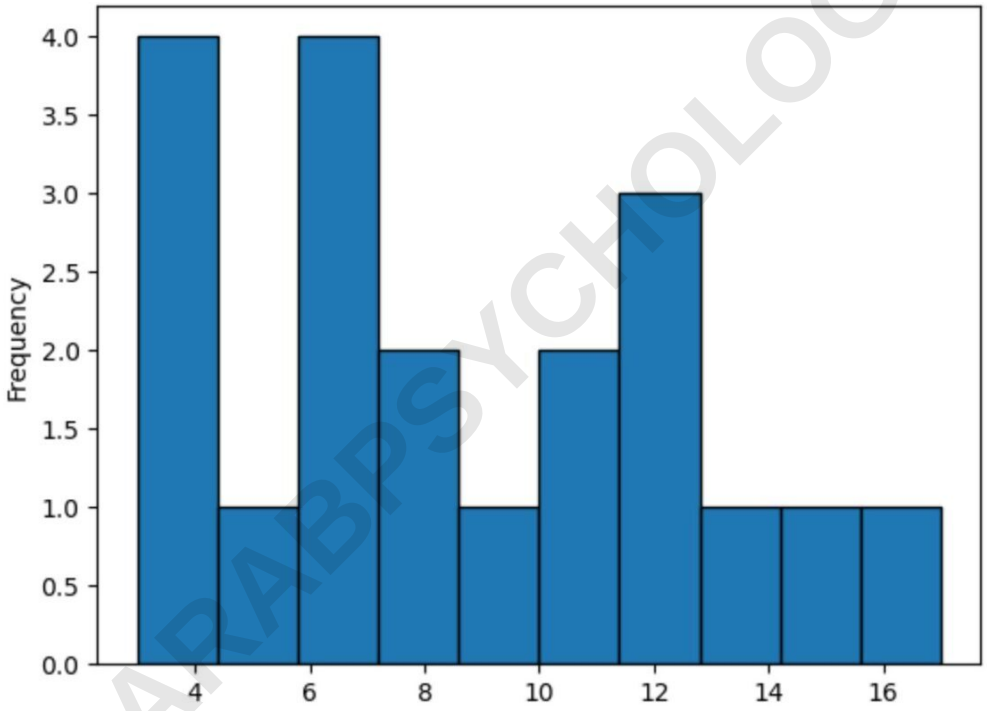


Note that kind='kde' tells pandas to use kernel density estimation, which produces a smooth curve that

summarizes the distribution of values for a variable.

If you'd like to create a histogram instead, you can specify kind='hist' as follows:

#plot distribution of values in points column using histogram
df.plot(kind='hist', edgecolor='black')



This method uses bars to represent frequencies of values in the points column as opposed to a smooth line that summarizes the shape of the distribution.

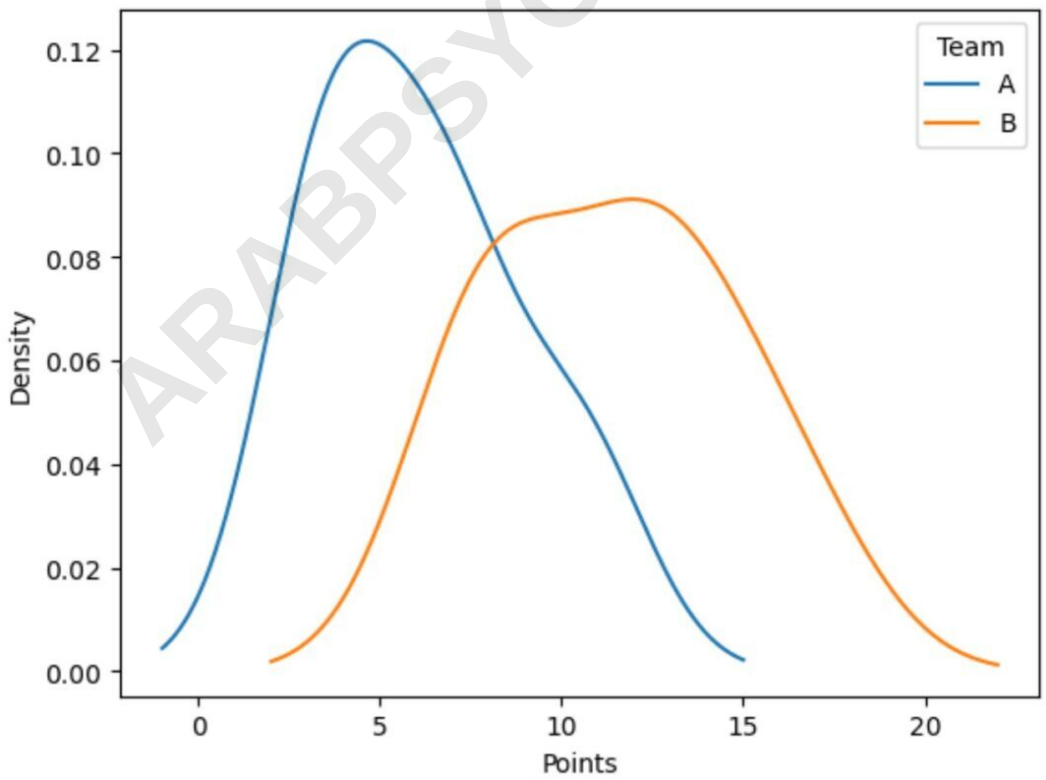
Example 2: Plot Distribution of Values in One Column, Grouped by Another Column

```
import matplotlib.pyplot as plt

#plot distribution of points by team
df.groupby('team').plot(kind='kde')

#add legend
plt.legend(, title='Team')

#add x-axis label
plt.xlabel('Points')
```



The blue line shows the distribution of points for players on team A while the orange line shows the distribution of points for players on team B.

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