

Chapter 2 :



Informatics

Practices

**Class XII (As per
CBSE Board)**

An illustration of a laptop computer with a white body and a black keyboard. The screen is open and displays the text "Python pandas-Histogram & Quantiles" in red. The laptop is set against a background of orange binary code (0s and 1s).

**Python pandas-
Histogram &
Quantiles**

A purple starburst graphic with multiple points, containing the text "New Syllabus 2019-20" in blue.

**New
Syllabus
2019-20**

Visit : python.mykvs.in for regular updates

Histogram

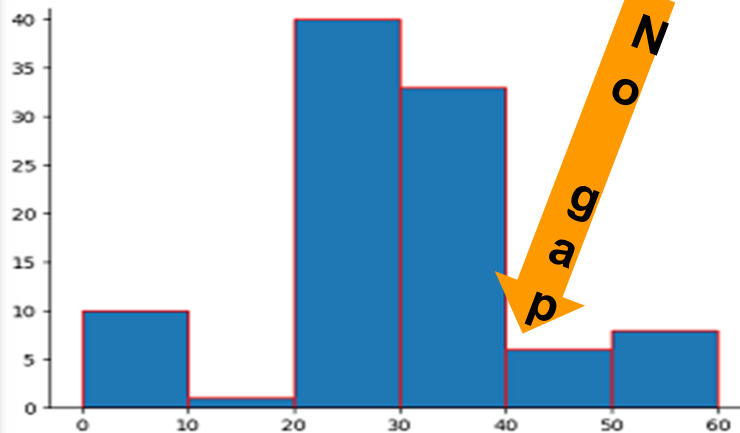
A histogram is a powerful technique in data visualization. It is an accurate graphical representation of the distribution of numerical data. It was first introduced by Karl Pearson. It is an estimate of the distribution of a continuous variable (quantitative variable). It is similar to bar graph. To construct a histogram, the first step is to “bin” the range of values—means divide the entire range of values into a series of intervals—and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.

Histogram

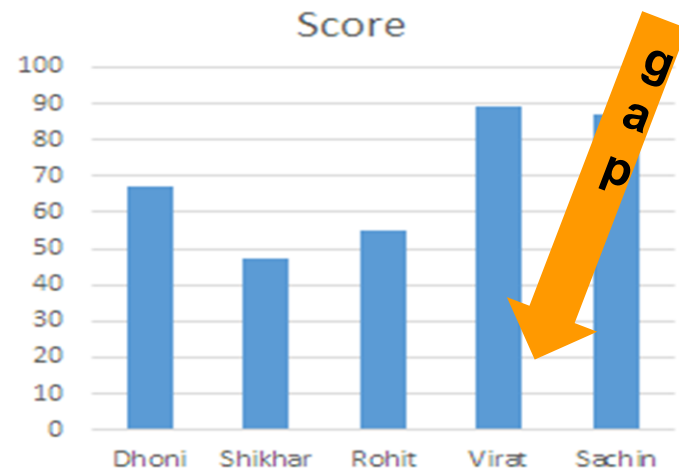
Difference between a histogram and a bar chart / graph –

A bar chart majorly represents categorical data (data that has some labels associated with it), they are usually represented using rectangular bars with lengths proportional to the values that they represent.

While histograms on the other hand, is used to describe distributions. Given a set of data, what are their distributions



HISTOGRAM



BAR CHART

Histogram

Histogram in Python –

Drawing a histogram in Python is very easy. All we have to do is code for 3-4 lines of code. But complexity is involved when we are trying to deal with live data for visualization.

To draw histogram in python following concepts must be clear.

Title –To display heading of the histogram.

Color – To show the color of the bar.

Axis: y-axis and x-axis.

Data: The data can be represented as an array.

Height and width of bars. This is determined based on the analysis. The width of the bar is called bin or intervals.

Border color –To display border color of the bar.

Histogram

Histogram in Python –

There are various ways to create histogram in python pandas. One of them is using matplotlib python library. Using this library we can easily create histogram. We have to write just few statements to create histogram.

So install matplotlib library using following statements at command prompt.

>pip install matplotlib

After installation we can create histogram. if pip does not work then copy the pip.exe file to the folder where we want to run the above command or move to the folder of pip.exe then write above command.

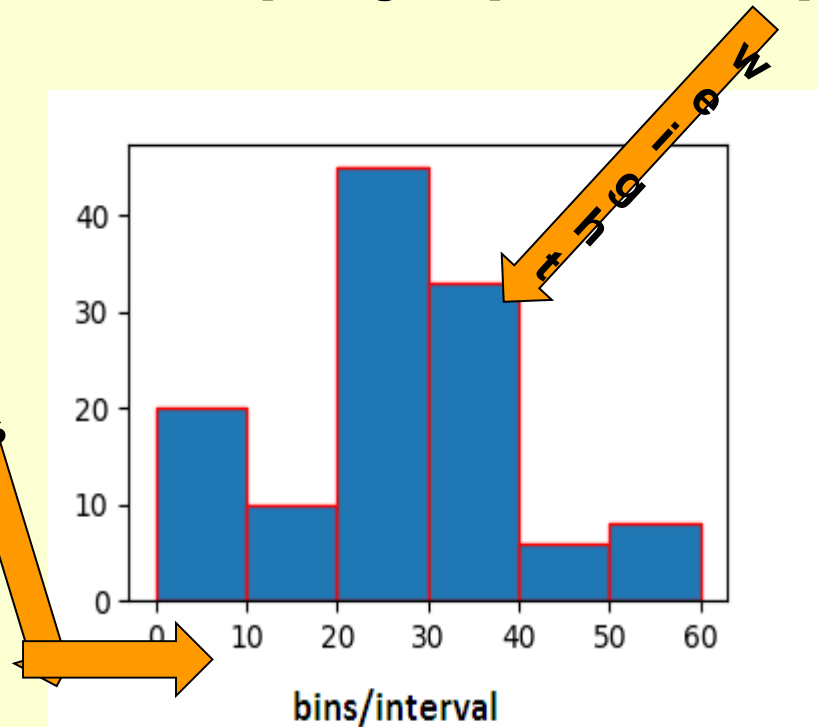
Histogram

Histogram in Python –

E.g. Program in python. Develop a python program with below code and execute it.

```
import numpy as np
import matplotlib.pyplot as plt
data = [1,11,21,31,41]
plt.hist([5,15,25,35,45, 55], bins=[0,10,20,30,40,50, 60], weights=[20,10,45,33,6,8],
edgecolor="red")
plt.show()
```

#first argument of hist() method is position (x,y Coordinate) of weight, where weight is to be displayed.
No of coordinates must match with No of weight otherwise error will generate
#Second argument is interval
#Third argument is weight for bars



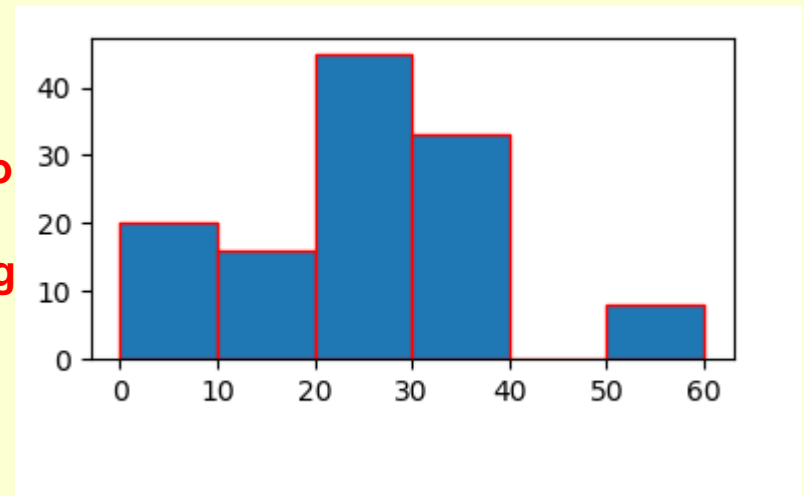
Histogram

Histogram in Python –

For better understanding we develop the same program with minor change .

```
import numpy as np
import matplotlib.pyplot as plt
data = [1,11,21,31,41]
plt.hist([5,15,25,35,15, 55], bins=[0,10,20,30,40,50, 60], weights=[20,10,45,33,6,8],
edgecolor="red")
plt.show()
```

at interval(bin)40 to 50 no bar because we have not mentioned position from 40 to 50 in first argument(list) of hist method. Where as in interval 10 to 20 width is being Displayed as 16 (10+6 both weights are added) because 15 is twice In first argument.



Histogram

Histogram in Python –

By default bars of histogram is displayed in blue color but we can change it to other color with following code .

```
plt.hist([1,11,21,31,41, 51], bins=[0,10,20,30,40,50, 60], weights=[10,1,0,33,6,8], facecolor='y', edgecolor="red")
```

In above code we are passing 'y' as facecolor means yellow color to be displayed in bars.

To give a name to the histogram write below code before calling show()

```
plt.title("Histogram Heading")
```

he histogram can be saved by clicking on the Save button on the GUI. Also, the following code will save the histogram as a PNG image.

```
plt.savefig("temp.png")
```

For x and y label below code can be written

```
plt.xlabel('Value')
```

```
plt.ylabel('Frequency')
```


Quantiles

The word “quantile” comes from the word quantity. means, a quantile is where a sample is divided into equal-sized or subgroups (that’s why it’s sometimes called a “fractile“). So that’s why ,It can also refer to dividing a probability distribution into areas of equal probability.

The median is a kind of quantile; the median is placed in a probability distribution at center so that exactly half of the data is lower than the median and half of the data is above the median. The median cuts a distribution into two equal parts and so why sometimes it is called 2-quantile.

Quartiles are quantiles; when they divide the distribution into four equal parts. Deciles are quantiles that divide a distribution into 10 equal parts and Percentiles when that divide a distribution into 100 equal parts .

Quantiles

How to Find Quantiles?

Sample question: Find the number in the following set of data where 30 percent of values fall below it, and 70 percent fall above:

2 4 5 7 9 11 12 17 19 21 22 31 35 36 45 44 55 68 79 80 81 88 90 91 92 100 112
113 114 120 121 132 145 148 149 152 157 170 180 190

Step 1: Order the data from smallest to largest. The data in the question is already in ascending order.

Step 2: Count how many observations you have in your data set. this particular data set has 40 items.

Step 3: Convert any percentage to a decimal for “q”. We are looking for the number where 30 percent of the values fall below it, so convert that to .3.

Step 4: Insert your values into the formula:

$$\text{ith observation} = q (n + 1)$$

$$\text{ith observation} = .3 (40 + 1) = 12.3$$

Answer: The ith observation is at 12.3, so we round down to 12 (remembering that this formula is an estimate). The 12th number in the set is 31, which is the number where 30 percent of the values fall below it.

Quantiles

How to Find Quantiles in python
In pandas series object->

```
import pandas as pd
import numpy as np
s = pd.Series([1, 2, 4, 5,6,8,10,12,16,20])
r=s.quantile(.3)
print(r)
```

OUTPUT

4.6999999999999999

Note – It returns 30% quantile

Quantiles

How to Find Quantiles in python
In pandas dataframe object->

```
import pandas as pd
import numpy as np
df = pd.DataFrame(np.array([[11, 1], [12, 10], [13, 100], [14, 100], [15, 1000]]),
                  columns=['a', 'b'])
r=df.quantile(.2)
print(r)
```

OUTPUT

a 11.8

b 8.2

Name: 0.2, dtype: float64

Note – It returns 20% quantile