Data Visualization

What is Data Visualization ?

Data visualization is the technique to present the data in a pictorial or graphical format. It enables stakeholders and decision makers to analyze data visually. The data in a graphical format allows them to identify new trends and patterns easily.

The main benefits of data visualization are as follows:

- ✓   It simplifies the complex quantitative information
- ✓   It helps analyze and explore big data easily
- ✓ It identifies the areas that need attention or improvement
- ✓ It identifies the relationship between data points and variables
- ✓   It explores new patterns and reveals hidden patterns in the data

Purpose of Data visualization:

- ➢ Better analysis
- ➢ Quick action
- ➢ Identifying patterns
- ➢ Finding errors
- ➢ Understanding the story
- ➢ Exploring business insights
- ➢ Grasping the Latest Trends

matplotlib Library and pyplot Interface
• The matplotlib is a python library that provides many interfaces functionally for 2D graphics
• In short we can call mattplotlib as a high quality plotting library of Python.
• The matplotlib library offers many different named collections of methods, pyplot is one such interface.
• pyplot is a collection of methods within matplotlib which allows user to construct

2D plots easily and interactively.

Installing matplotlib

It is done using pip command in **Command Prompt**

  pip install matplotlib

Importing PyPlot

To import Pyplot following syntax is
        import matplotlib.pyplot
             or
        import matplotlib.pyplot as plt

After importing matplotlib in the form of **plt** we can use **plt** for accessing any function of matplotlib

Steps to plot in matplotlib:

- Create a .py file & import matplotlib library to it using import statement

        import matplotlib.pyplot as plt

- Set data points in plot( ) method of plt object
- Customize plot by setting different parameters

- Call the show() method to display the plot

- Save the plot/graph if required

Types of plot using matplotlib

- **LINE PLOT**

- **BAR GRAPH**

- **HISTOGRAM etc.**

Line Plot:

A line plot/chart is a graph that shows the frequency of data occurring along a number line. The line plot is represented by a series of data points called **markers** connected with a straight line. Generally, line plots are used to display trends over time. A line plot or line graph can be created using the plot () function available in pyplot library.

We can, not only just plot a line but we can explicitly define the grid, the x and y axis scale and labels, title and display options etc.

Line chart: displaying data in form of lines.

- We can create line graph with x coordinate only or with x and y coordinates.
- Function to draw line chart – plot()
- Default colour of line- blue

- The default **width** for each bar is .**0.8** units, which can be changed.

• Syntax: plt.plot(x,y)

Line Plot customization
   • **Custom line color**
   plt.plot(x,y,'red')
   Change the value in color argument like 'b' for blue,'r','c',.....

   • **Custom line style and line width**
   plt.plot(x,y, linestyle='solid' , linewidth=4).
   set linestyle to solid/dashed/dotted/dashdot
   set linewidth as required
   • **Title**
   plt.title('DAY – TEMP Graph ') – Change it as per requirement

   • **Label**-
           plt.xlabel('TIme') – to set the x axis label
            plt.ylabel('Temp') – to set the y axis label
- Changing Marker Type, Size and Color
           plt.plot(x,y,'blue',marker='*',markersize=10,markeredgecolor='magenta')

 Order of methods used in plot() function:
 Plt.plot(x,y,color,linewidth,linestyle,marker, markersize,markeredgecolor)

Function used to show the graph – show()
       plt.show( )
PROGRAM
   ```
   import matplotlib.pyplot as plt
   X=[1,2,3,4,5]
   Y=[2,4,6,8,10]
   plt.title('Simple Line Graph')
   plt.xlabel('X Axis')
   plt.ylabel('Y Axis')
   plt.plot(X,Y,'r')
   plt.show()
   ```
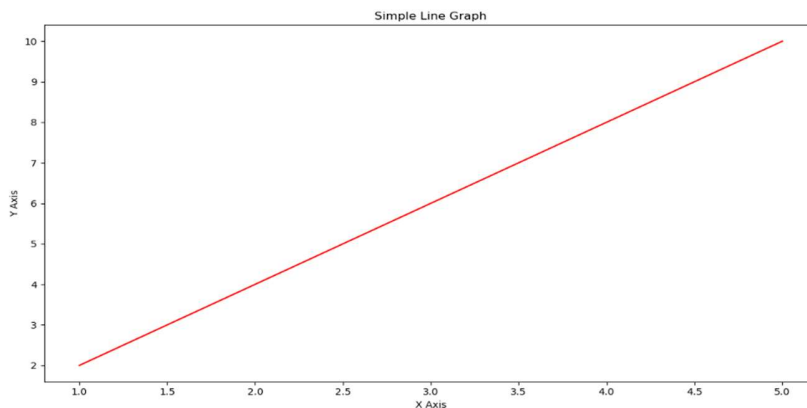
Bar Graph

A graph drawn using rectangular bars to show how large each value is. The bars can be horizontal or vertical. A bar graph makes it easy to compare data between different groups at a glance. Bar graph represents categories on one axis and a discrete value in the other. The goal of bar graph is to show the relationship between the two axes. Bar graph can also show big changes in data over time.

- Syntax :   plt.bar(x,y)

Bar graph customization

• **Custom bar color**

    plt.bar(x,y, color="color code/color name")

To se different colors for different bars

    plt.bar(x,y, color="color code/color name sequence")

• **Custom bar width**

    plt.bar(x,y, width=float value)

To set different widths for different bars

    plt.bar(x,y, width=float value sequence)

• **Title**

plt.title(' Bar Graph ') – Change it as per requirement
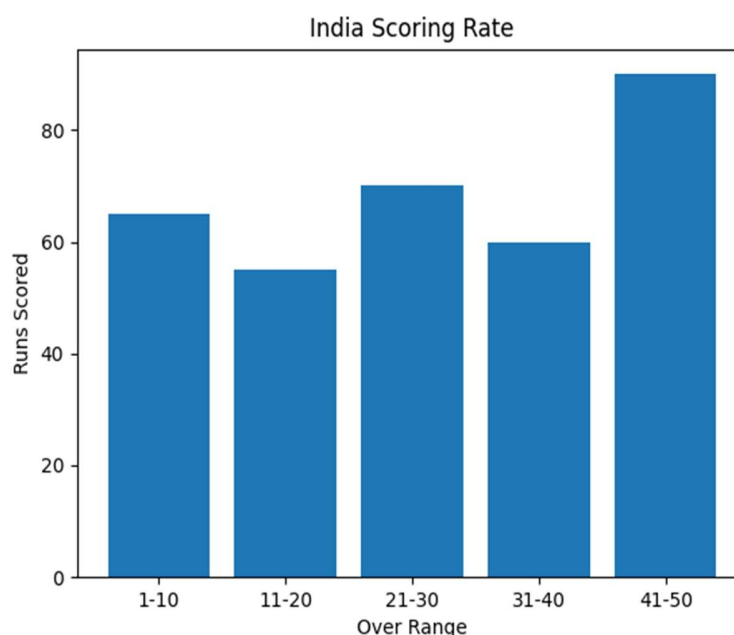
• **Label-**

            plt.xlabel('Overs') – to set the x axis label

            plt.ylabel('Runs') – to set the y axis label

PROGRAM :

```
import matplotlib.pyplot as plt
overs=['1-10','11-20','21-30','31-40','41-50']
runs=[65,55,70,60,90]
plt.xlabel('Over Range')
plt.ylabel('Runs Scored')
plt.title('India Scoring Rate')
plt.bar(overs,runs)
plt.show( )
```
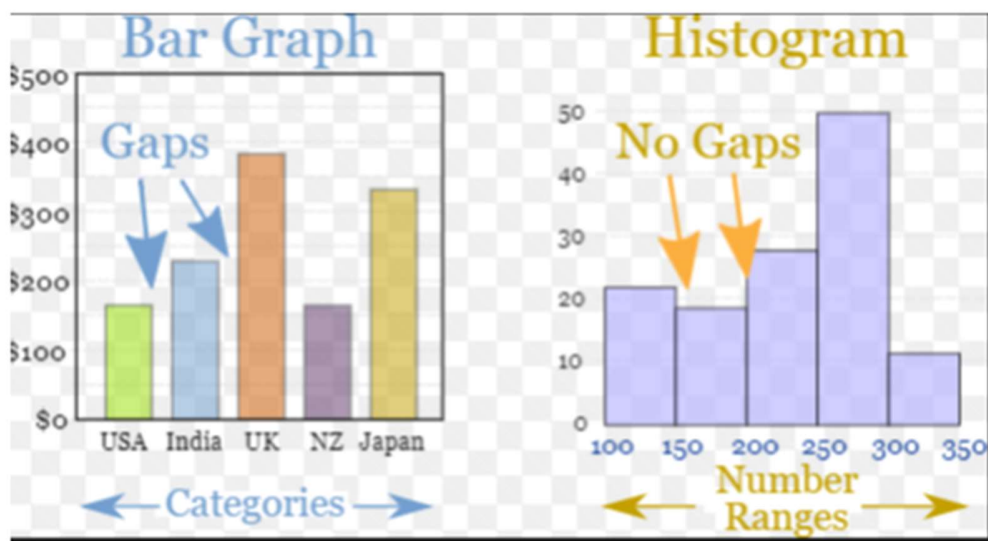
HISTOGRAM

A histogram is a graphical representation which organizes a group of data points into user specified ranges.

Histogram provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values ("bins"). It is similar to a vertical bar graph but without gaps between the bars.

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.

Creating a Histogram :

- It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.

- To create a histogram the first step is to create bin of the ranges, then distribute the whole range of the values into a series of intervals, and count the values which fall into each of the intervals.

- Bins are clearly identified as consecutive, non-overlapping intervals of variables.

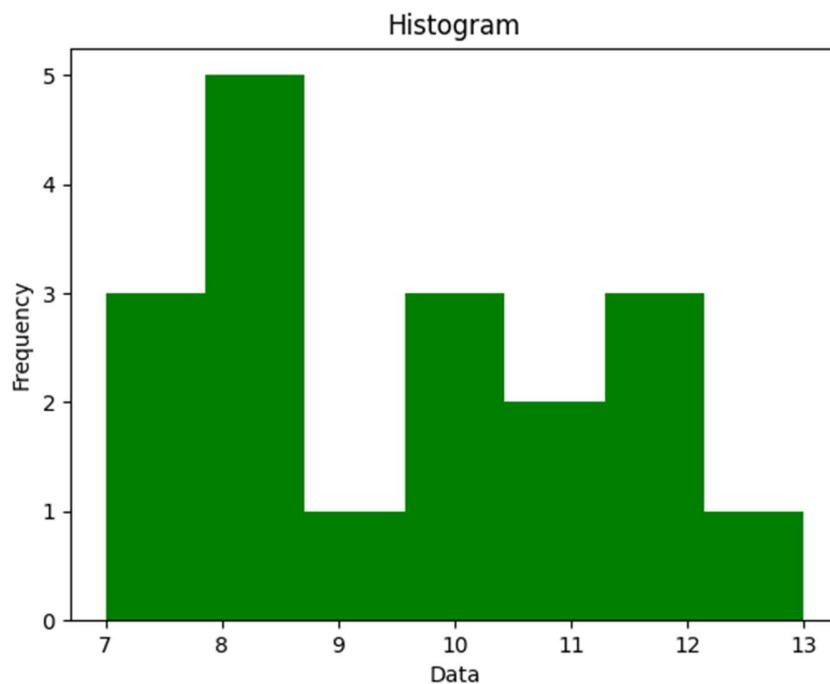- The hist() function is used to create histogram

- **Syntax:**

  plt.hist(x,other parameters)

Optional Parameters

x	array or sequence of array
bins	optional parameter contains integer or sequence or strings
histtype	optional parameter used to create type of histogram [bar, barstacked, step, stepfilled], default is "bar"
align	optional parameter controls the plotting of histogram [left, right, mid]
orientation	Optional. Possible values are 'horizontal' or 'vertical'
color	optional parameter used to set color or sequence of color specs

PROGRAM :

```
import matplotlib.pyplot as plt
data=[7,7,7,8,8,8,8,8,9,10,10,10,11,11,12,12,12,13]
plt.xlabel('Data')
plt.ylabel('Frequency')
plt.title('Histogram')
plt.hist(data,bins=7,color='green')
plt.show()
```

• **Title**

plt.title('Histogram ') – Change it as per requirement

• **Label**-

plt.xlabel('Data') – to set the x axis label

plt.ylabel('Frequency') – to set the y axis label

• **Legend -** A legend is an area describing the elements of the graph. In the matplotlib library there is a function named legend() which is used to place a legend on the axes .
When we plot multiple ranges in a single plot ,it becomes necessary that legends are specified.It is a color or mark linked to a specific data range plotted .

To plot a legend you need to do two things.

i)In the plotting function like bar() or plot() , give a specific label to the data range using label

ii)Add legend to the plot using legend ( ) as per the sytax given below .

   Syntax : - plt.legend((loc=position number or string)

   position number **can be** u1,2,3,4 specifying the position strings upper right/'upper left/'lower left/lower right respectively .

60

Default position is upper right or 1

Saving the Plot

To save any plot savefig() method is used. Plots can be saved in various formats like pdf,png,eps etc .

plt.savefig('line_plot.pdf') // save plot in the current directory

plt.savefig('d:\\plot\\line_plot.pdf')   // save plot in the given path

WORKSHEET 1

1.What is data visualization?

a) It is the numerical representation of information and data

b) It is the graphical representation of information and data

c) It is the character representation of information and data

 d) None of the above

2.Which is a python package used for 2D graphics?

 a) matplotlib.pyplot

 b) matplotlib.pip

c) matplotlib.numpy

 d) mathplotlib.pyplot

3.The command used to give a heading to a graph is _____

(a) plt.show()

(b) plt.plot()

(c) plt.xlabel()

 (d) plt.title()

4. Using Python Matplotlib _____ can be used to count how many values fall into each interval.

 (a) line plot

(b) bar graph

(c) histogram

(d) None of these

5.Fill the missing statement

import matplotlib.pyplot as plt

 marks=[30,10,55,70,50,25,75,49,28,81]

 plt._____(marks, bins='auto', color='green')

plt.show()

(a) plot

 (b) bar

(c)hist

(d)draw

6.Which module of matplotlib library is required for plotting of graph ?
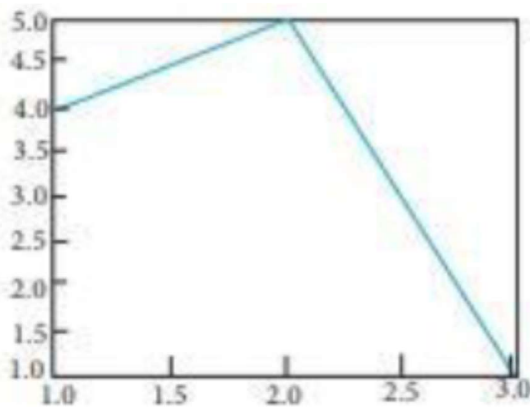
(a) Plot

(b) Matplot

(c) pyplot

(d) graphics

7.Observe the output figure. Identify the code for obtaining this output.

a)   import matplotlib.pyplot as plt
     plt.plot([1,2],[4,5])
     plt.show()
b)   import matplotlib.pyplot as plt
     plt.plot([2,3],[5,1])
     plt.show()
c)   import matplotlib.pyplot as plt
     plt.plot([1,2,3],[4,5,1])
     plt.show()
d)   import matplotlib.pyplot as plt

```
plt.plot([1,3],[4,1])
plt.show()
```

8.Identify the right type of chart using the following hints.
Hint 1: This chart is often used to visualize a trend in data over intervals of time.
Hint 2: The line in this type of chart is often drawn chronologically.
a) Line chart
b) Bar chart
c) Pie chart
d) Scatter plot

9.Which of the following is/are correct statement for plot method?
a) plt.plot(x,y,color,others)
b) pl.plot(x,y)
c) pl.plot(x,y,color)
d) All the above

10. Consider the following graph. Write the code to plot it

ANSWERS

1. **b) It is the graphical representation of information and data**
2. **a) matplotlib.pyplot**
3. **d) plt.title()**
4. **c) histogram**
5. **c)hist**
6. **c) pyplot**
7. **c) import matplotlib.pyplot as plt**
   plt.plot([1,2,3],[4,5,1])
   plt.show()

8. **a) Line chart**
9. **d) All the above**
10. **import matplotlib.pyplot as plt**
    a = [1,2,3,4,5]
    b = [10,31,26,24,20]
    plt.plot(a,b)
    plt.show()

WORKSHEET 2

1.To give a title to x-axis, which of the following
method is used?
a) plt.xtitle("title")
b) plt.xlabel("title")
c) plt.xheader("title")
d) plt.xlabel.show("title")

2.To change the width of bars in bar chart, which of the following argument
with a float value is used?
a) thick
b) thickness
c) width
d) barwidth

3.What is the purpose of legend?
a) A legend is an area describing the elements of the graph.
b) A legend is top area with information about graph
c) A legend is additional information of x and y labels
d) A legend is a mini box with bars data

4.Which function can be used to export generated graph in matplotlib to
png
1.   a) savefigure ( )
2.   b) savefig( )
3.   c) save( )
4.   d) export ( )

5.which one of these is not a valid line style in matplotlib
a) '-'
b) '--'
c) '-.'
d) '<'

6.How can we make bar chart horizontal?
a) plt.bar()
b) plt.hbar()
c) plt.barh()
d) plt.rightbar()

7. A histogram is used:
a) for continuous data
b) for grouped data
c) for time series data
d) to compare two sets of data
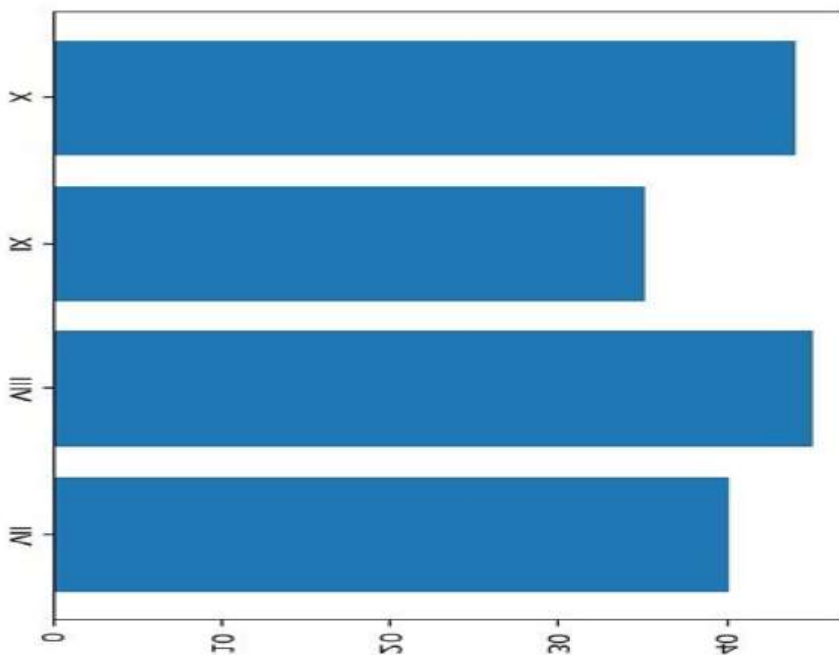8.Which function is used to show legend ?
a) display ( )
b) show( )
c) legend( )
d) legends( )

9.The datapoints plotted on a graph are called _____
a) Markers
b) Values
c) Ticks
d) Pointers

10.Write code to draw the following bar graph representing the classes and number of students in each class.

Answers:

1. **b) plt.xlabel("title")**
2. **c) width**
3. **a) A legend is an area describing the elements of the graph.**
4. **b) savefig( )**
5. **d) '<'**
6. **c) plt.barh()**
7. **a) for continuous data**
8. **c) legend( )**
9. **a)Markers**
10.
   import matplotlib.pyplot as plt
   classes = ['VII','VIII','IX','X']
   students = [40,45,35,44]
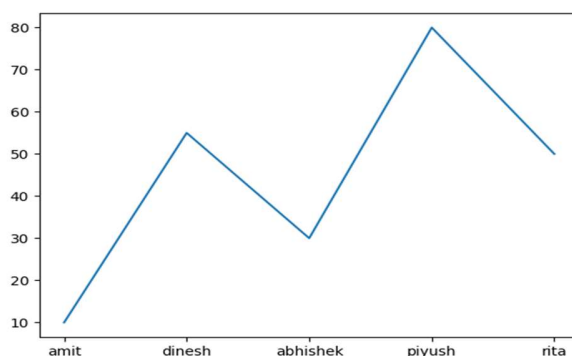   **plt.barh(classes, students)**
   plt.show()

WORKSHEET 3

1.To specify the style of line as dashed , which argument of plot() needs to be set ?

a) line
b) width
c) Style
d) linestyle

2. Which of the following is not a valid plotting function in pyplot?
   a) bar()
   b) hist()
   c) histh()
   d) barh()

3.Observe the following figure. Identify the coding for obtaining this as output.

a) import matplotlib.pyplot as plt

eng_marks=[10,55,30,80,50]
st_name=["amit","dinesh","abhishek","piyush","rita"]
plt.plot(st_name,eng_marks)
plt.show()

b) import matplotlib.pyplot as plt
eng_marks=[10,55,30,80,50]
st_name=["amit","dinesh","abhishek","piyush","rita"]
plt.plot(st_name,eng_marks)

c) import matplotlib.pyplot as plt
eng_marks=[10,55,30,80,50]
st_name=["amit","dinesh","abhishek","piyush","rita"]
plt.plot(eng_marks, st_name)
plt.show()

d) import matplotlib.pyplot as plt
eng_marks=[10,55,30,80,50]
st_name=["amit","dinesh","abhishek","piyush","rita"]
plt.plot(eng_marks, st_name)
plt.show()

4.Read the statements given below and identify the right option to draw a histogram.

Statement A: To make a Histogram with Matplotlib, we can use the plt.hist() function.

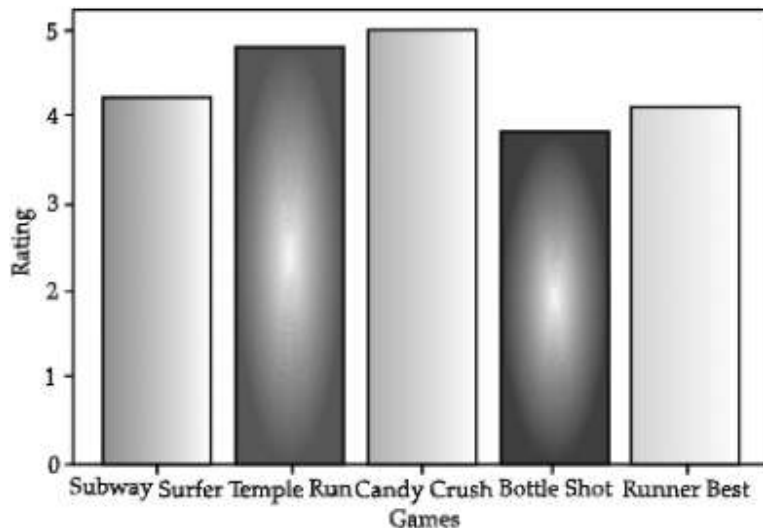**Statement B:** The bin parameter is compulsory to create histogram.

   a) Statement A is correct
   b) Statement B is correct
   c) Statement A is correct, but Statement B is incorrect
   d) d. Statement A is incorrect, but Statement B is correct

5. Which graph should be used where each column represents a range of values, and the height of a column corresponds to how many values are in that range?
   a) plot
   b) line
   c) bar
   d) histogram

Go through the following case study and answer the questions 6 to 10.

Mr. Sharma is working in a game development industry and he was comparing the given chart on the basis of the rating of the various games available on the play store. He is trying to write a code to plot the graph. Help Mr. Sharma to fill in the blanks of the code and get the desired output.

import_____ #Statement 1

Games=["Subway Surfer","Temple Run","Candy Crush","Bottle hot","Runner Best"]

Rating=[4.2,4.8,5.0,3.8,4.1]

plt._____(Games,Rating) #Statement 2

plt.xlabel("Games")

plt._____("Rating") #Statement 3

plt._____ #Statement 4

 6) Choose the right code from the following for statement 1.

(a) matplotlib as plt

(b) pyplot as plt

(c) matplotlib.pyplot as plt

(d) matplotlib.plt as pyplot

7) Identify the name of the function that should be used in statement 2 to plot the above graph.

 (a) line()

(b) bar()

(c) hist()

d) barh()

8) Choose the correct option for the statement 3.

(a) title("Rating")

(b) ytitle("Rating")

(c) ylabel("Rating")

(d) yaxis("Rating")

9) Choose the right function/method from the following for the statement 4.

(a) display()

(b) print()

(c) bar()

(d) show()

10) In case Mr. Sharma wants to change the above plot to any other shape, which statement, should he change.

 (a) Statement 1

(b) Statement 2

(c) Statement 3

(d) Statement 4

11.Write a program to generate a histogram with the following values

X=[25,28,35,46,57,68,73]

Include the following parameters/arguments along with historam

1.It should be a horizontal histogram
2.Number of bins should be 20

<p align="center">ANSWERS</p>

1. d) linestyle
2. c)histh( )

3. import matplotlib.pyplot as plt
    eng_marks=[10,55,30,80,50]
    st_name=["amit","dinesh","abhishek","piyush","rita"]
    plt.plot(st_name,eng_marks)
    plt.show()
4. c) Statement A is correct, but Statement B is incorrect
5. d). histogram
6. c) matplotlib.pyplot as plt
7. b) bar()
8. c) ylabel("Rating")
9. d) show()
10. b) Statement 2
11.

```
import matplotlib.pyplot as plt
X=[25,28,35,46,57,68,73]
plt.hist(X,orientation='horizontal',bins=20)
plt.show()
```