UNIT 5:
Data Literacy –
Data Collection to Data Analysis

Jnanknana Koshti

Learning Objectives:

1. To understand the importance of data literacy in AI.
2. To explore various data collection methods and their applications.
3. To analyse data using basic Statistical analysis techniques.
4. To identify matrices and their role in representing data like images.
5. To understand the preparation of data to suit the models.

Demystifying Data Types:

Before diving into projects, establish a strong foundation by exploring data types:

- Nominal, Ordinal, Interval, Ratio: Unveil the different scales of measurement for data points. Nominal (category labels), Ordinal (ranked order), Interval (consistent units but no true zero), and Ratio (true zero allows meaningful comparisons).
- Quantitative vs. Qualitative: Distinguish between data that is numerical (quantitative) and data that describes characteristics (qualitative).
- Discrete vs. Continuous: Help students differentiate between data with distinct values (discrete) and data that can take any value within a range (continuous).
- Structured, Semi-structured, Unstructured: Introduce the various data organization formats. Structured data follows a predefined schema, semi-structured allows some flexibility, and unstructured data has no fixed format.

The Cyclical Nature of Data Projects:

Emphasize the iterative nature of data projects. Data is used at every stage of the cycle:

- Problem Definition: Data informs the problem you're trying to solve.
- Data Collection: Gathering the right data is crucial for analysis.
- Data Cleaning and Preparation: Data is processed to ensure accuracy and usability.
- Data Analysis and Exploration: Data is analysed to uncover patterns and insights.
- Visualization and Communication: Findings are communicated through visualizations.

Essential Libraries:

- Statistics Library: Introduce the 'statistics' library for performing common statistical calculations on data (e.g., mean, median, standard deviation).

- Matplotlib Library: Showcase the 'matplotlib' library for creating various data visualizations (e.g., histograms, scatter plots) to represent data insights effectively.

1. WHAT IS DATA LITERACY?

- Think about all the information you encounter daily – online, in books, from friends. Can you categorize this information in any way? Is it all the same?

- Imagine you're trying to solve a mystery. What kind of clues would you need to gather and analyse to figure things out?

- Data can be defined as a representation of facts or instructions about some entity (students, school, sports, business, animals etc.) that can be processed or communicated by human or machines.

- It is a widely known fact that Artificial Intelligence (AI) is essentially data-driven. AI involves converting large amounts of raw data into actionable information that carry practical value and is usable.

- Data literacy means being able to find and use data effectively.

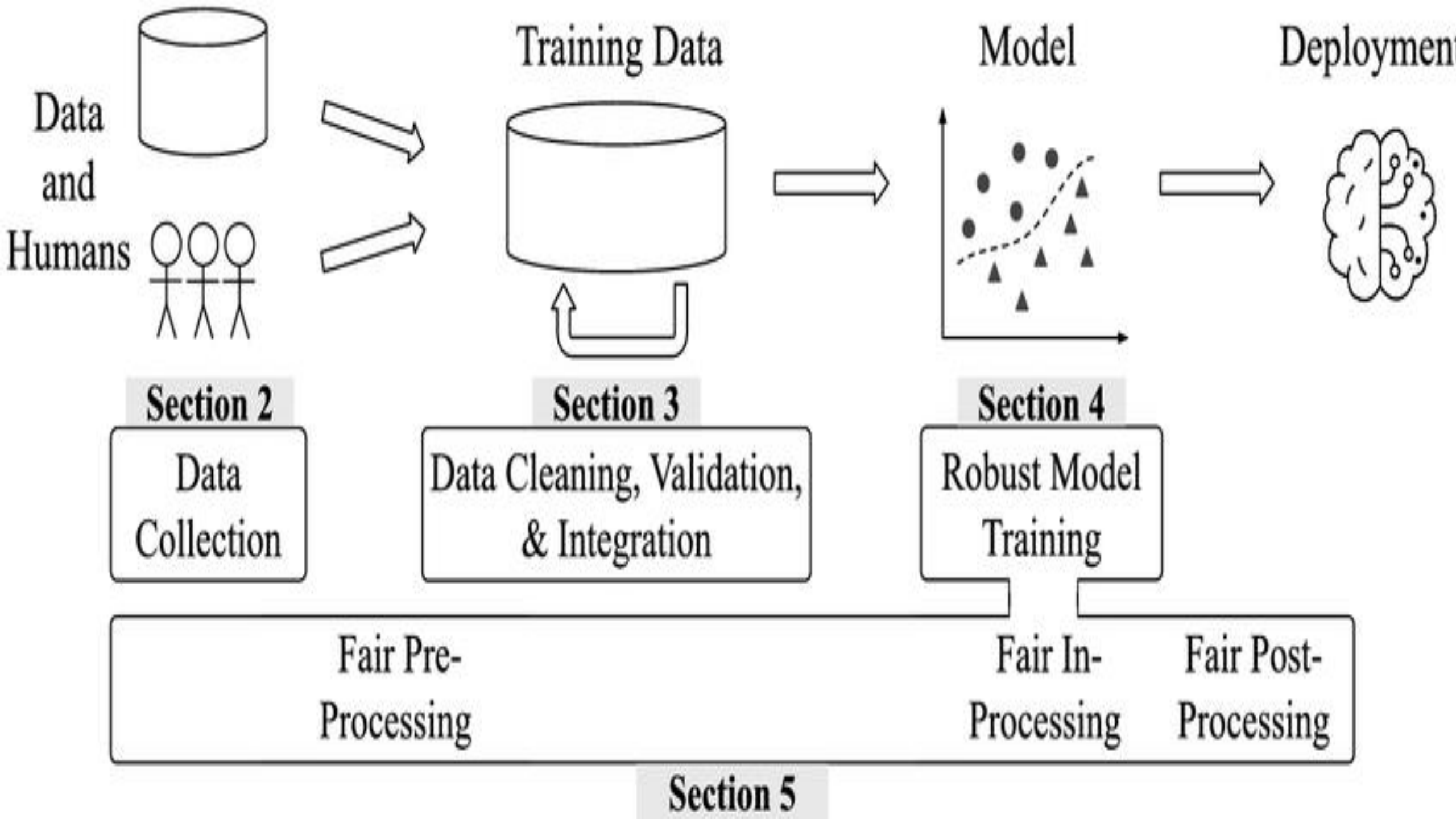
- This includes skills like collecting data, organizing it, checking its quality, analysing it, understanding the results and using it ethically.

- Data may be structured, semi structured or unstructured.
- It should be collected, organized and analysed properly to know whether the input for AI models is valid and appropriate or not.
- AI data analysis involves using AI techniques and data science to improve the processes of cleaning, inspecting, and modelling both structured and unstructured data.
- The primary objective is to extract valuable information that can support decision-making and drawing conclusions.

2. DATA COLLECTION

- Think about your favourite movie recommendation platform. How do you think they use data to suggest movies you might like?

- Imagine you wanted to build a robot that could sort recycling. What kind of data would you need to collect to train it to recognize different materials?


Data
and
Humans
Training Data
Model
Deployment
Section 2
Section 3
Section 4
Data
Collection
Data Cleaning, Validation,
& Integration
Robust Model
Training
Fair Pre-
Processing
Fair In-
Processing
Fair Post-
Processing
Section 5

- Data collection allows you to capture a record of past events so that we can use data analysis to find recurring patterns.

- Data collection means pooling data by scraping, capturing, and loading it from multiple sources, including offline and online sources.

- It is recommended to collect as much data as possible for good predictions.
- The most important thing to consider while data collection is diversity.
- Diverse data will help your model cover more scenarios.
- The quantity of data also depends on the complexity of your model.
- If it is as simple as license plate detection then you can expect predictions with small batches of data.
- But if you are working on higher levels of Artificial intelligence like medical AI, you need to consider huge volumes of data.

- Data is the main ingredient of any Project.
- Hence the process of identifying the data requirements, its collection and analysis will be done iteratively.
- There are mainly two sources of data collection: Primary and Secondary.
- Primary Sources are sources which are created to collect the data for analysis.

Method	Description	Example
Survey	Gathering data from a population through interviews, questionnaires, or online forms. Useful for measuring opinions, behaviors, and demographics.	A researcher uses a questionnaire to understand consumer preferences for a new product.
Interview	Direct communication with individuals or groups to gather information. It can be structured, semi-structured, or unstructured.	An organization conducts an online survey to collect employee feedback about job satisfaction.
Observation	Watching and recording behaviors or events as they occur. Often used in ethnographic research or when direct interaction is not possible.	Observing children's play patterns in a schoolyard to understand social dynamics.
Experiment	Manipulating variables to observe their effects on outcomes. Used to establish cause-and-effect relationships.	Testing the effectiveness of different advertising campaigns on a group of people.

Marketing Campaign (using data)	Utilizing customer data to predict behavior and optimize campaign performance.	A company personalizes email marketing campaigns based on past customer purchases.
Questionnaire	A specific tool used within surveys - a list of questions designed to gather data from respondents. You can collect quantitative (numerical) or qualitative (descriptive) information.	A questionnaire might ask respondents to rate their satisfaction on a scale of 1 to 5 and also provide open-ended feedback.

- Secondary data sources are where the data is already stored and ready for use.

- Data given in Books, journals, News Papers, Websites, Internal transactional databases, etc can be reused for data analysis.

Method	Description	Example
Social Media Data Tracking	Collecting data from social media platforms like user posts, comments, and interactions.	Analyzing social media sentiment to understand audience reception towards a new product launch.
Web Scraping	Using automated tools to extract specific content and data from websites.	Scraping product information and prices from e-commerce websites for price comparison.
Satellite Data Tracking	Gathering information about the Earth's surface and atmosphere using satellites.	Monitoring weather patterns and environmental changes using satellite imagery.
Online Data Platforms	Websites offering pre-compiled datasets for various purposes.	Kaggle, GitHub etc.

3. EXPLORING DATA

- Imagine you're collecting data on student’s favourite movie genres. Could you rank the genres from most to least popular (ordinal)? Or would you just be able to say which genre is the favourite (nominal)?

- When measuring temperature, we can say it's 20 degrees Celsius today, which is 10 degrees warmer than yesterday. Can we say it's twice as hot today (interval)? Why or why not?

Data Exploration

Initial Phase of Data Analysis

Involves

Data

Cleaning

Visualizing

Summarizing

to

Uncover Patterns

Anomalies

Insights

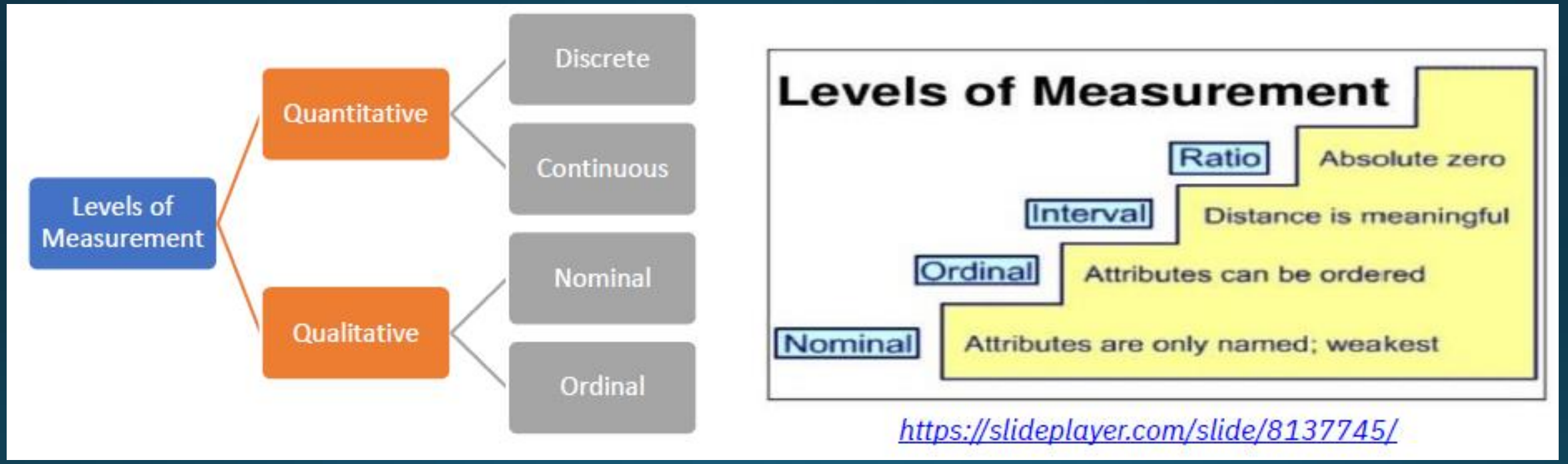
for

Further Analysis

EDUCBA

- Exploring data is about "getting to know" the data: and its values - whether they are typical, unusual, spread out, or whether they are extremes.

- Levels of Measurement



https://slideplayer.com/slide/8137745/

Types of Data

- Data can be classified according to level of measurement.
- The level of measurement dictates the calculations that can be done to summarize and present the data.
- It also determines the statistical tests that should be performed.

Level of measurement

Nominal

Data may only be classified

- Jersey numbers of football player.
- Make of car.

Ordinal

Data are ranked no meaningful difference between values

- Your rank in class.
- Team standings.

Interval

Meaningful difference between values.

- Temperature
- Dress size

Ratio

Meaningful 0 point and ratio between values.

- No. of patients seen
- No of sales call made
- Distance students travel to class

1. Nominal Level

- Nominal variables are like categories such as Mercedes, BMW or Audi, or like the four seasons – winter, spring, summer and autumn.
- They aren’t numbers, and cannot be used in calculations and neither in any order or rank.
- The nominal level of measurement is the simplest or lowest of the four ways to characterize data.
- Nominal means "in name only".

Example 1:

Please indicate your marital status.

____ Married ____ Single ____ Separated ____ Divorced ____ Widowed

---

Example 2:

Do you like or dislike chocolate ice cream?

____ Like ____ Dislike

Example 3:

Which of the following supermarkets have you shopped at in the last 30 days? Please check all that apply.

____ Albertson's ____ Winn-Dixie ____ Publix ____ Safeway ____ Walmart

---

Example 4:

Please indicate your gender:

____ Female ____ Male ____ Transgender

2. Ordinal Level

- Ordinal data, is made up of groups and categories which follow a strict order.

- For e.g. if you have been asked to rate a meal at a restaurant and the options are: unpalatable, unappetizing, just okay, tasty, and delicious.

Example 1:

How likely are you to recommend the Santa Fe Grill to a friend?	Definitely Will Not Recommend						Definitely Will Recommend
	1	2	3	4	5	6	7

Example 2:

Using a scale of 0–10, with "10" being Highly Satisfied and "0" being Not Satisfied At All, how satisfied are you with the banking services you currently receive from (read name of primary bank)?

Answer: _____

Example 3:

Please indicate how frequently you use different banking methods. For each of the banking methods listed below, circle the number that best describes the frequency you typically use each method.

Banking Methods	**Never Use**									**Use Very Often**
Inside the bank	0	1	2	3	4	5	6	7	8	9	10
Drive-up window	0	1	2	3	4	5	6	7	8	9	10
24-hour ATM	0	1	2	3	4	5	6	7	8	9	10
Debit card	0	1	2	3	4	5	6	7	8	9	10
Bank by mail	0	1	2	3	4	5	6	7	8	9	10
Bank by phone	0	1	2	3	4	5	6	7	8	9	10
Bank by Internet	0	1	2	3	4	5	6	7	8	9	10

3. Interval Level

- Data that is measured using the interval scale is similar to ordinal level data because it has a definite ordering but there is a difference between the two data.

- The differences between interval scale data can be measured though the data does not have a starting point i.e. zero value.

- Temperature scales like Celsius (° C) and Fahrenheit (° F) are measured by using the interval scale.

Interval level:

- One category is higher than another (Ordered).
- There is a constant unit of measurement.
- Zero is just a point on the scale; or there is no natural zero point.
- Division of two numbers does not make sense.
- Scale or rank are good examples

- EXAMPLE: Temperature on the Fahrenheit scale.
  - Zero is just a point on the scale.
- EXAMPLE: Shoe size and dress size.
  - There is no natural zero point
- EXAMPLE: Years in which Whole Foods Market Inc. stock split.
  - Division of 1992 and 1993 does not make sense.
- EXAMPLES: Rank of Indi 500 results, Test scores.

4. Ratio Scale Level

- Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated.

- For example, the scores of four multiple choice statistics final exam questions were recorded as 80, 68, 20 and 92 (out of a maximum of 100 marks).

Examples

- **Discrete ratio data**
  - Number of children in a household
  - Number of vehicles owned in a specific period (5 years)
  - Number of male students in a classroom
- **Continuous ratio data**
  - Years of working experience
  - Number of hours spent in a waiting room
  - Driving speed (Mph)

Activity-1

Student Health Survey – Fill in the response and mention appropriate Level of Measurement.

Query	Response	Level of Measurement
Sex (Male/ Female)		
Height (in metres)		
Weight (in kilograms)		
Rate overall health (Excellent; Good; Average; Below Average; Poor)		
Pulse rate (in BPM)		
Body temperature (in Fahrenheit)		
Country of residence		

Activity-2.

Indicate whether the variable is ordinal or not. Write the variable type, if it is not ordinal.

- Opinion about a new law (favour or oppose)

_____

- Letter grade in an English class (A, B, C, etc.)

_____

- Student rating of teacher on a scale of 1 – 10.

_____

4. STATISTICAL ANALYSIS OF DATA

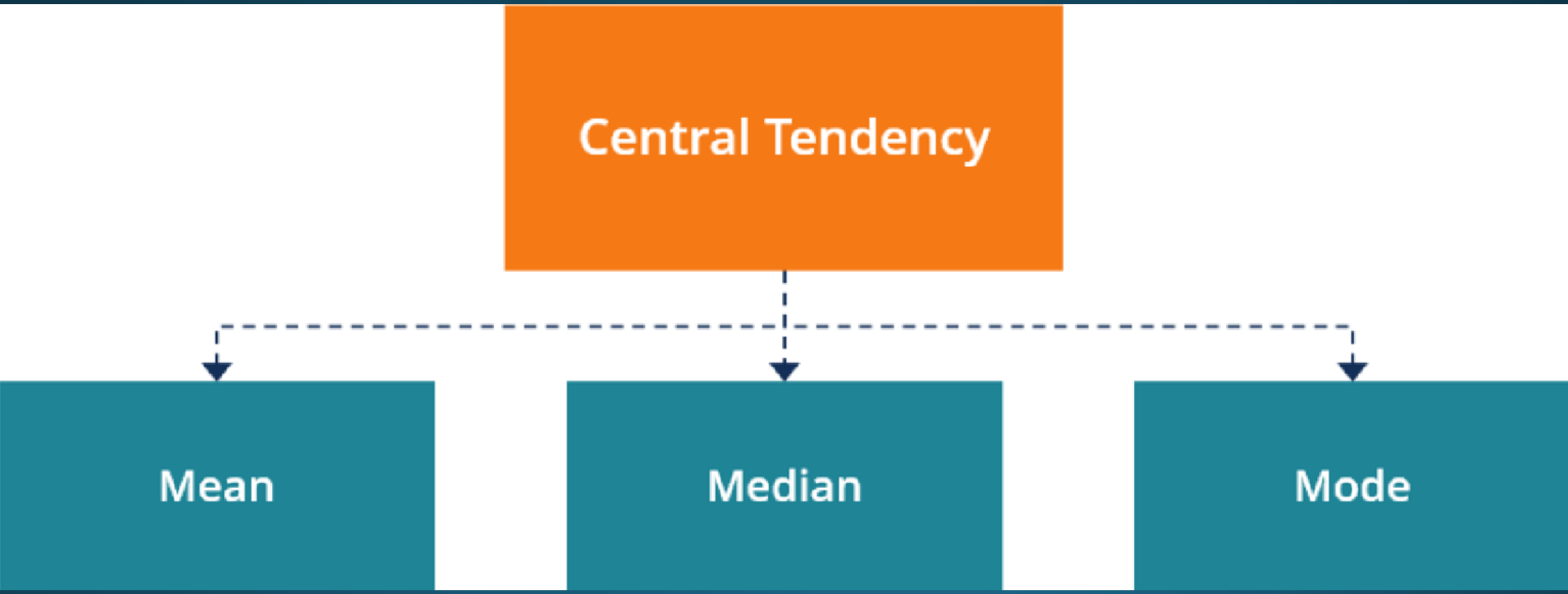
- Imagine you have a dataset of the heights of all students in your class. How would you find the "average" height? Is there just one way, or could there be different ways to measure it depending on the data?

- We often hear about data being "spread out" or "clumped together." How can we describe how spread out the data in a set is, besides just knowing the average value?

Measure of Central Tendency

- Statistics is the science of data, which is in fact a collection of mathematical techniques that helps to extract information from data.

- For the AI perspective, statistics transforms observations into information that you can understand and share.


Central Tendency
Mean
Median
Mode

- “Central tendency” is stated as the summary of a data set in a single value that represents the entire distribution of data domain (or data set).

- We can perform Statistical Analysis using Python programming language. For that we have to import the library statistics into the Program

- mean ( ) →returns the mean of the data
- median ( ) →returns the median of the data
- mode ( ) →returns the mode of the data
- variance ( ) →returns the variance of the data
- stdev ( ) →returns the standard deviation of the data

Mean

- In statistics, the mean (more technically the arithmetic mean or sample mean) can be estimated from a sample of examples drawn from the domain. It is a quotient obtained by dividing the total of the values of a variable by the total number of their observations or items.

$$M = \Sigma fx / n$$

- where M = Mean
- Σ = Sum total of the scores
- f = Frequency of the distribution
- x = Scores
- n = Total number of cases

Example -1

The set S = {5,10,15,20,30}
Mean of set S = 5+10+15+20+30/5 = 80/5 = 16

Example- 2 Calculate the mean of the following grouped data.

Class	Frequency
2 - 4	3
4 - 6	4
6 – 8	2
8 – 10	1

Program-1

- There are 25 students in a class. Their heights are given below.

- Write a Python Program to find the mean. heights → 145, 151, 152, 149, 147, 152, 151,149, 152, 151, 147, 148, 155, 147,152,151, 149,145, 147, 152,146, 148, 150, 152, 151

```
import statistics
height = [145,151, 152, 149, 147, 152, 151,149,
          152, 151, 147, 148, 155, 147,152,151,
          149,145, 147, 152,146, 148, 150, 152, 151]
print ("Mean height of students", statistics.mean(height))
```

OUTPUT

```
Mean height of students 149.56
```

Median

- The median is another measure of central tendency. It is positional value of the variables which divides the group into two equal parts, one part comprising all values greater than median and other part smaller than median.

Example-3

Following series shows marks in mathematics of students learning AI

17	32	35	15	21	41	32	11	10	20	27	28	30

We arrange this data in an ascending or descending order.

10, 11, 15, 17, 20, 21, 27, 28, 30, 32, 32, 35, 40

As 27 is in the middle of this data position wise, therefore

Median = 27

Program-2

- There are 25 students in a class. Their heights are given below. Write a Python Program to find the median.
- heights → 145, 151, 152, 149, 147, 152, 151,149, 152, 151, 147, 148, 155, 147,152,151, 149,145, 147, 152,146, 148, 150, 152, 151

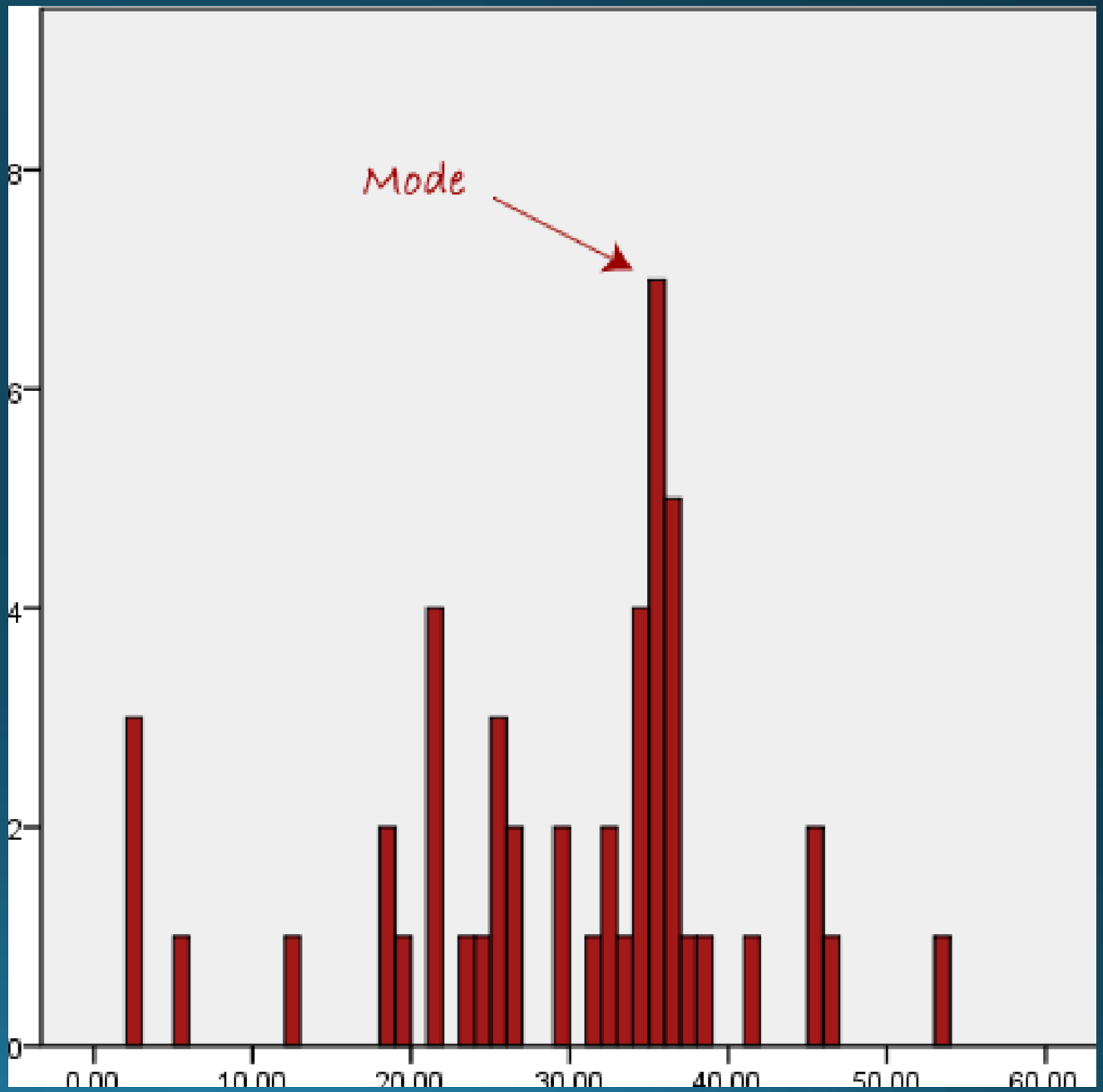
```
import statistics
height = [145,151, 152, 149, 147, 152, 151,149,
          152, 151, 147, 148, 155, 147,152,151,
          149,145, 147, 152,146, 148, 150, 152, 151]
print ("Median of height of students", statistics.median(height))
```

OUTPUT

```
Median of height of students 150
```

- Mode

Mode is another important measure of central tendency of statistical series. It is the value which occurs most frequently in the data series. It represents the highest bar in a bar chart or histogram. An example of a mode is presented below:



Example-4 Age of 15 students of a class
Age (years) 22, 24, 17, 18, 17, 19, 18, 21, 20, 21, 20, 23, 22, 22, 22,22,21,24

We arrange this series in ascending order as 17,17,18,18,19,20,20,21,21,22,22,22,

•

An inspection of the series shows that 22 occurs most frequently, hence
Mode=22

Program – 3

- Write a program to find the mode (heights → 145,151, 152, 149, 147, 152, 151,149, 147, 152, 151, 147, 148, 155, 147,152,151, 149, 145, 147, 152,146, 148, 150, 152, 151)

```
import statistics
height = [145,151, 152, 149, 147, 152, 151,149,
          152, 151, 147, 148, 155, 147,152,151,
          149,145, 147, 152,146, 148, 150, 152, 151]
print ("Mode of height of students", statistics.mode(height))
```

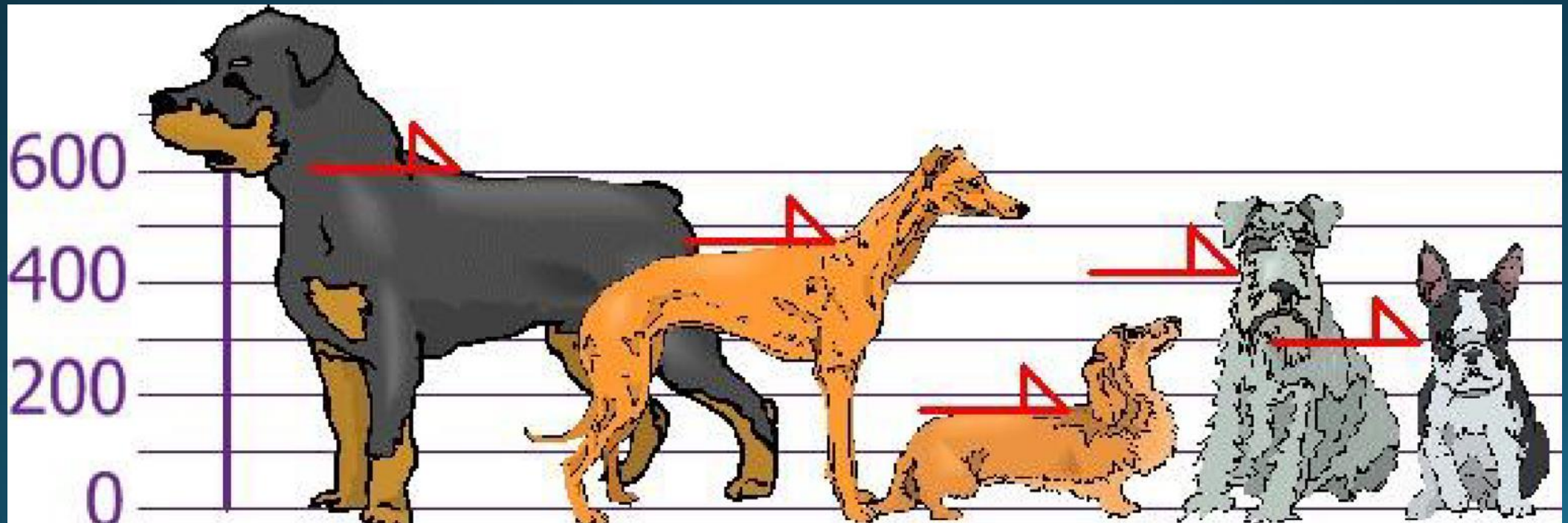
OUTPUT

```
Mode of height of students 152
```

Mean	Median	Mode
• The mean is a good measure of the central tendency when a data set contains values that are relatively evenly spread with no exceptionally high or low values.	• The median is a good measure of the central value when the data include exceptionally high or low values. The median is the most suitable measure of average for data classified on an ordinal scale.	• Mode is used when you need to find the distribution peak and peak may be many. For example, it is important to print more of the most popular books; because printing different books in equal numbers would cause a shortage of some books and an oversupply of others.

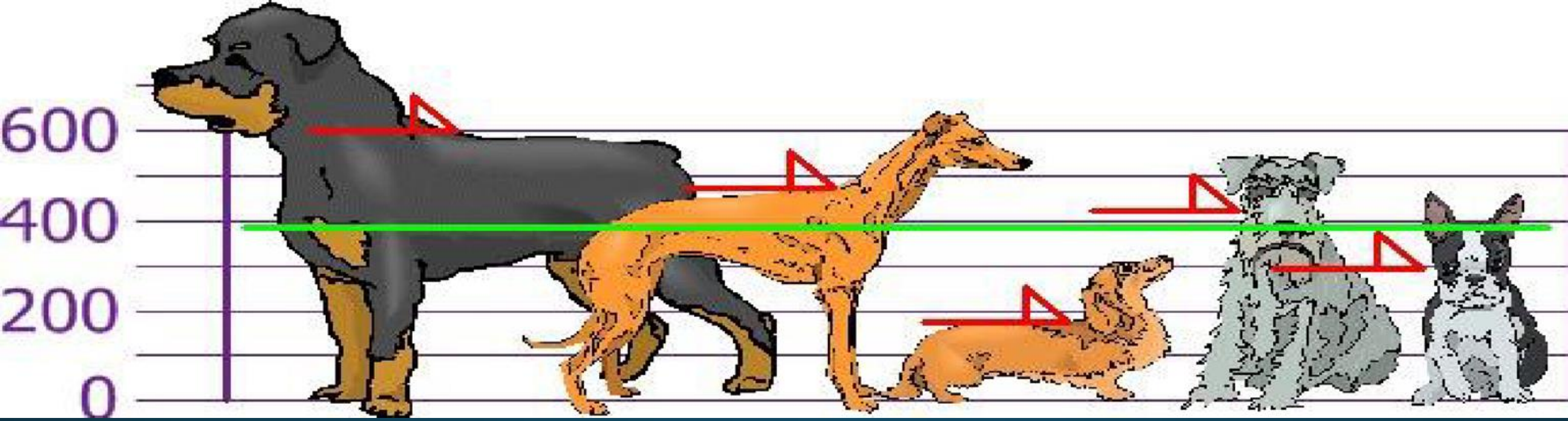
Variance and Standard Deviation

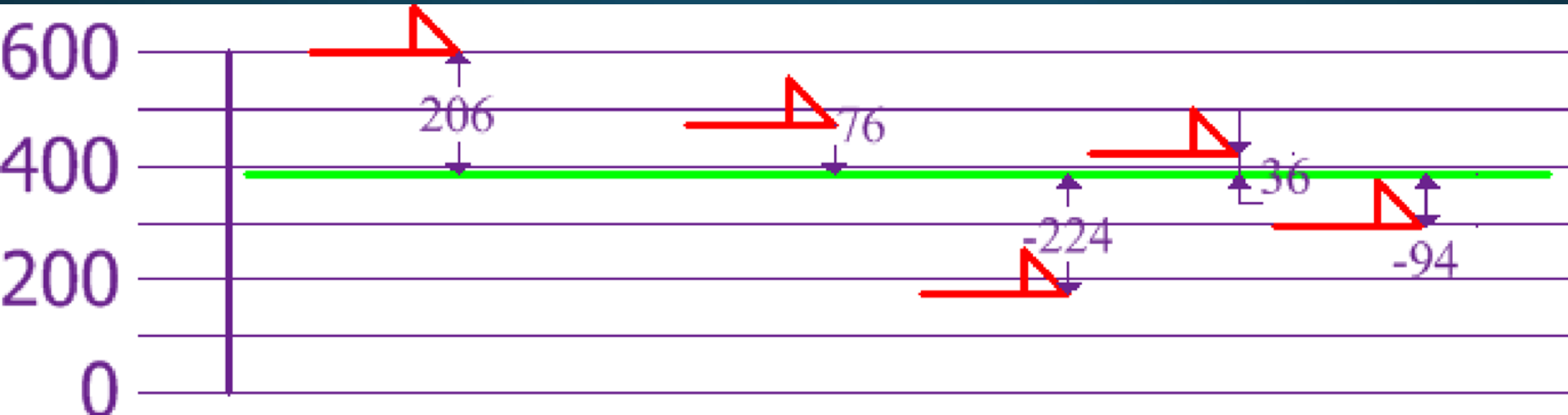
Measures of central tendency (mean, median and mode) provide the central value of the data set. Variance and standard deviation are the measures of dispersion (quartiles, percentiles, ranges), they provide information on the spread of the data around the centre.



As you can see, their heights are:
600mm,
470mm,
170mm,
430mm and
300mm.

Let us calculate their mean,
Mean = (600 + 470 + 170 + 430 + 300) / 5 = 1970 / 5 = 394 mm


600
400
200
0


600
400
200
0
206
76
-224
36
-94

- Calculate the difference (from mean height), square them, and find the average. This average is the value of the variance.

- Variance = [(206) 2 + (76) 2 + (-224) 2 + (36) 2 + (-94) 2] / 5 = 108520 / 5 = 21704

- And standard deviation is the square root of the variance.

- Standard deviation = $\sqrt{21704}$ = 147.32

FORMULA

Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Some important facts about variance and standard deviation.

- A small variance indicates that the data points tend to be very close to the mean, and to each other.
- A high variance indicates that the data points are very spread out from the mean, and from one another.
- A low standard deviation indicates that the data points tend to be very close to the mean.
- A high standard deviation indicates that the data points are spread out over a large range of values.

Program -4

- Write a program to find the variance and standard deviation. (heights → 145,151, 152, 149, 147, 152, 151,149, 152, 151, 147, 148, 155, 147,152,151, 149,145, 147, 152,146, 148, 150, 152, 151)

```
import statistics
height = [145,151, 152, 149, 147, 152, 151,149,
          152, 151, 147, 148, 155, 147,152,151,
          149,145, 147, 152,146, 148, 150, 152, 151]
print ("Variance in the height of students", statistics.variance(height))
print ("Standard Deviation", statistics.stdev(height))
```

OUTPUT

```
Variance in the height of students 6.756666666666667
Standard Deviation 2.599358953175872
```

5. REPRESENTATION OF DATA

- Imagine you have a big bag of colourful candies. How would you easily describe the different colours and how many candies there are of each colour without counting them all one by one?

- Have you ever seen a weather map or a graph in a book? Why do you think people use pictures and charts instead of just writing out numbers?

Data representation techniques are broadly classified in two ways:

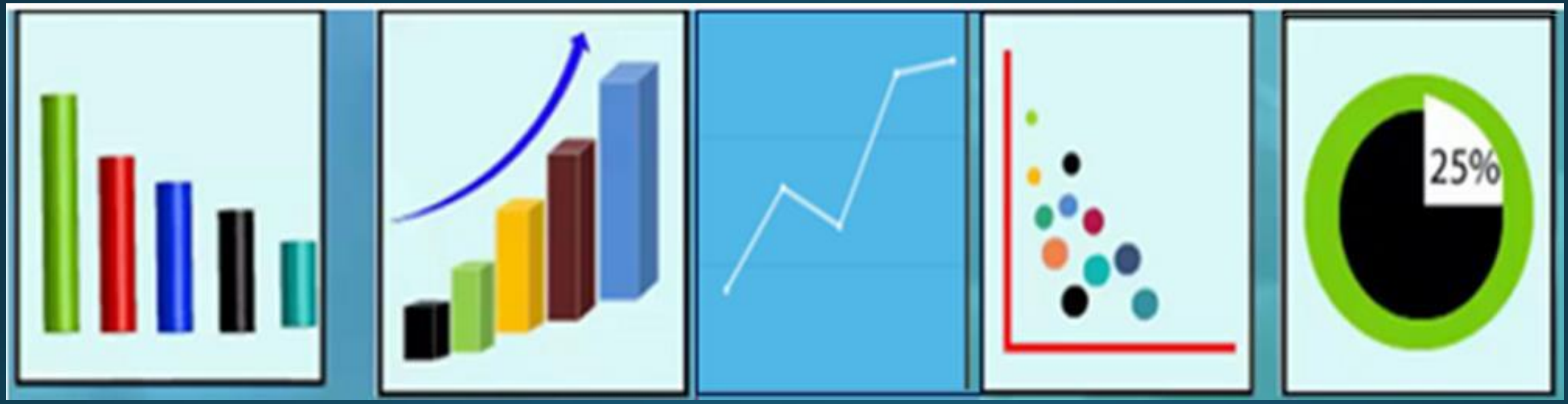
Non-Graphical technique:

Tabular form and case form: This is the old format of data representation not suitable for large datasets. Non-graphical techniques are not so suitable when our objective is to make some decisions after analysing a set of data.

Graphical Technique:

- The visual display of statistical data in the form of points, lines, dots and other geometrical forms is most common. For a complex and large quantity of data, human brain is more comfortable in dealing if represented through visual format means Graphical or pictorial representation of the data using graph, chart, etc. is known as Data visualization.

- Line graphs • Bar diagrams • Pie diagram
- Scatter Plots • Histogram



Data Visualization is possible in python using the library **Matplotlib**.

- pyplot is a submodule of Matplotlib that provides a MATLAB-like interface to the library. pyplot also provides a number of convenience functions that make it easy to create simple plots.

Installing Matplotlib

pip install matplotlib

or

python – m pip install – U matplotlib

In the program we have to import the library.

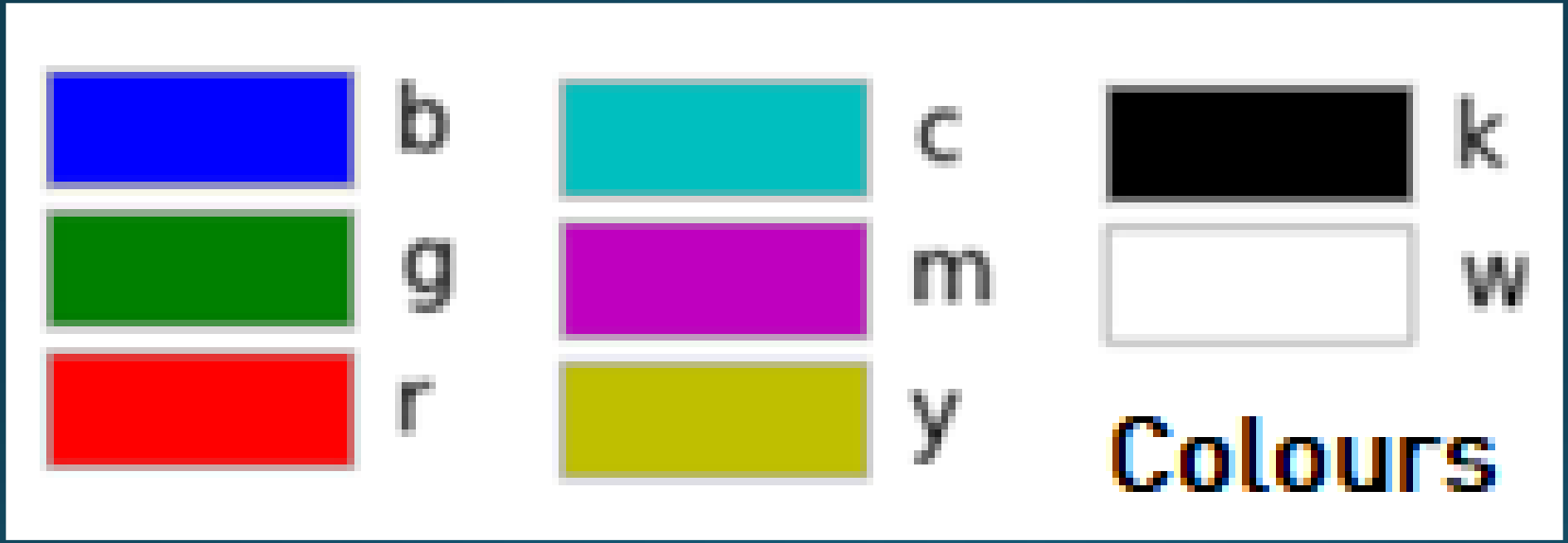
import matplotlib.pyplot

Function Name	Description
title ( )	Adds title to the chart/graph
xlabel ( )	Sets label for X-axis
ylabel ( )	Sets label for Y-axis
xlim ( )	Sets the value limit for X-axis
ylim( )	Sets the value limit for Y-axis
xticks ( )	Sets the tick marks in X-axis
yticks( )	Sets the tick marks in Y-axis
show ( )	Displays the graph in the screen
savefig("*adrress*")	Saves the graph in the address specified as argument.
figure (figsize = ***value in tuple format*****)**	Determines the size of the plot in which the graph is drawn. Values should be supplied in tuple format to the attribute figsize which is passed as argument.

List of Markers and its descriptions:

marker	symbol	description	marker	symbol	description
"."		point	"P"		plus (filled)
","		pixel	"*"		star
"o"		circle	"h"		hexagon1
"v"		triangle_down	"H"		hexagon2
"^"		triangle_up	"+"		plus
"s"		square	"x"		x
"p"		pentagon	"D"		diamond

List of Graph Colour Codes:



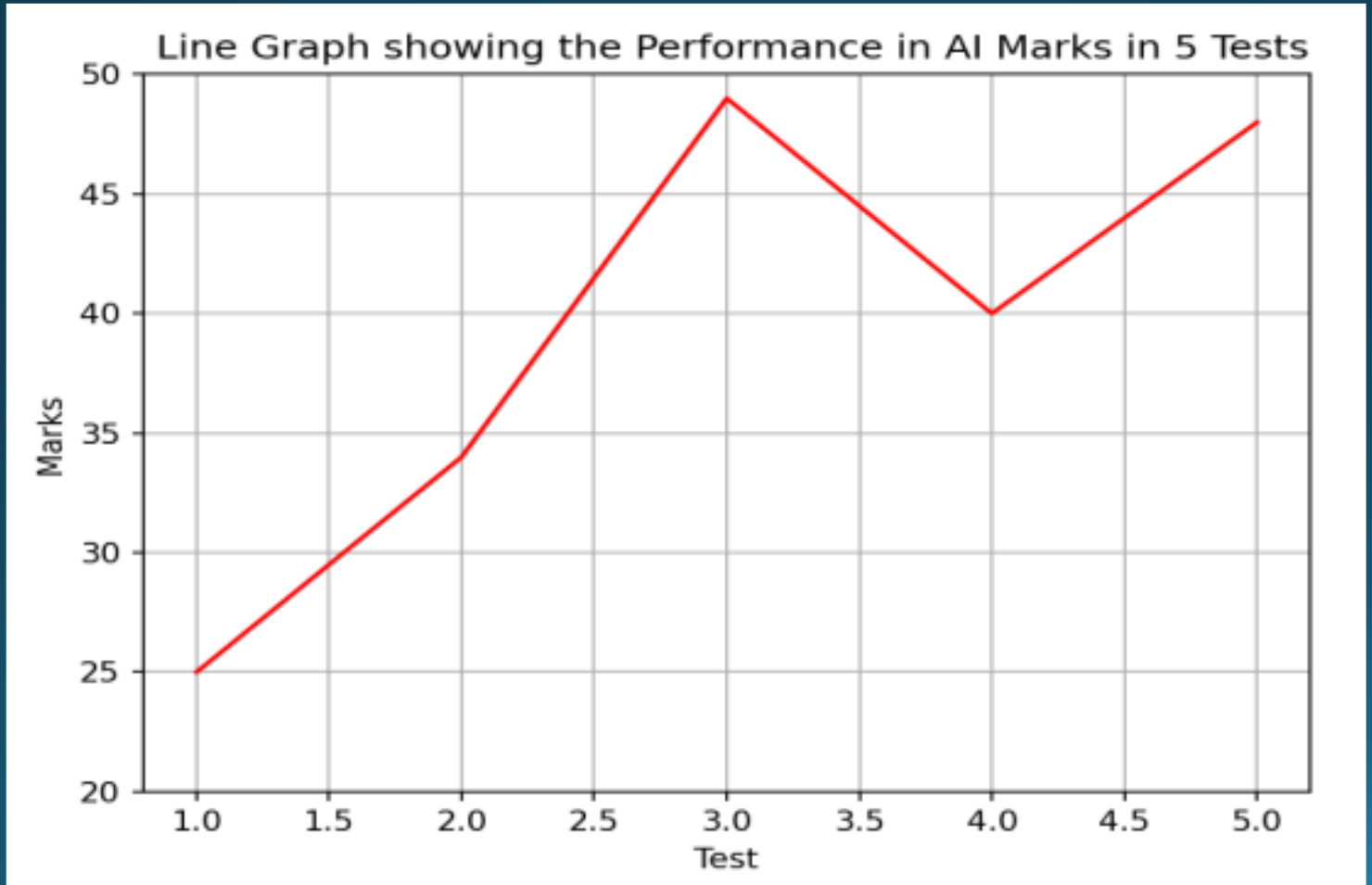
1. Line Graph

- A line graph is a powerful tool used to represent continuous data along a numbered axis.
- It allows us to visualize trends and changes in data points over time. Line graphs are suitable for data that can take on any value within a specific range.
- The line can slope upwards, indicating an increase, or downwards, signifying a decrease, reflecting the changes in the data over time.

Example-5:

- Kavya's AI marks for 5 consecutive tests is given below. Draw a line graph to Analyse her performance.

Test-1	Test-2	Test-3	Test-4	Test-5
25	34	49	40	48



Activity -3: Construct a simple line graph to represent the rainfall data of Kerala as shown in the table below

Month	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Rainfall (cm)	7.5	6.3	3.5	1.8	1.2	25.8	19.7	20.3	15.9	22.4	18.6	11.2

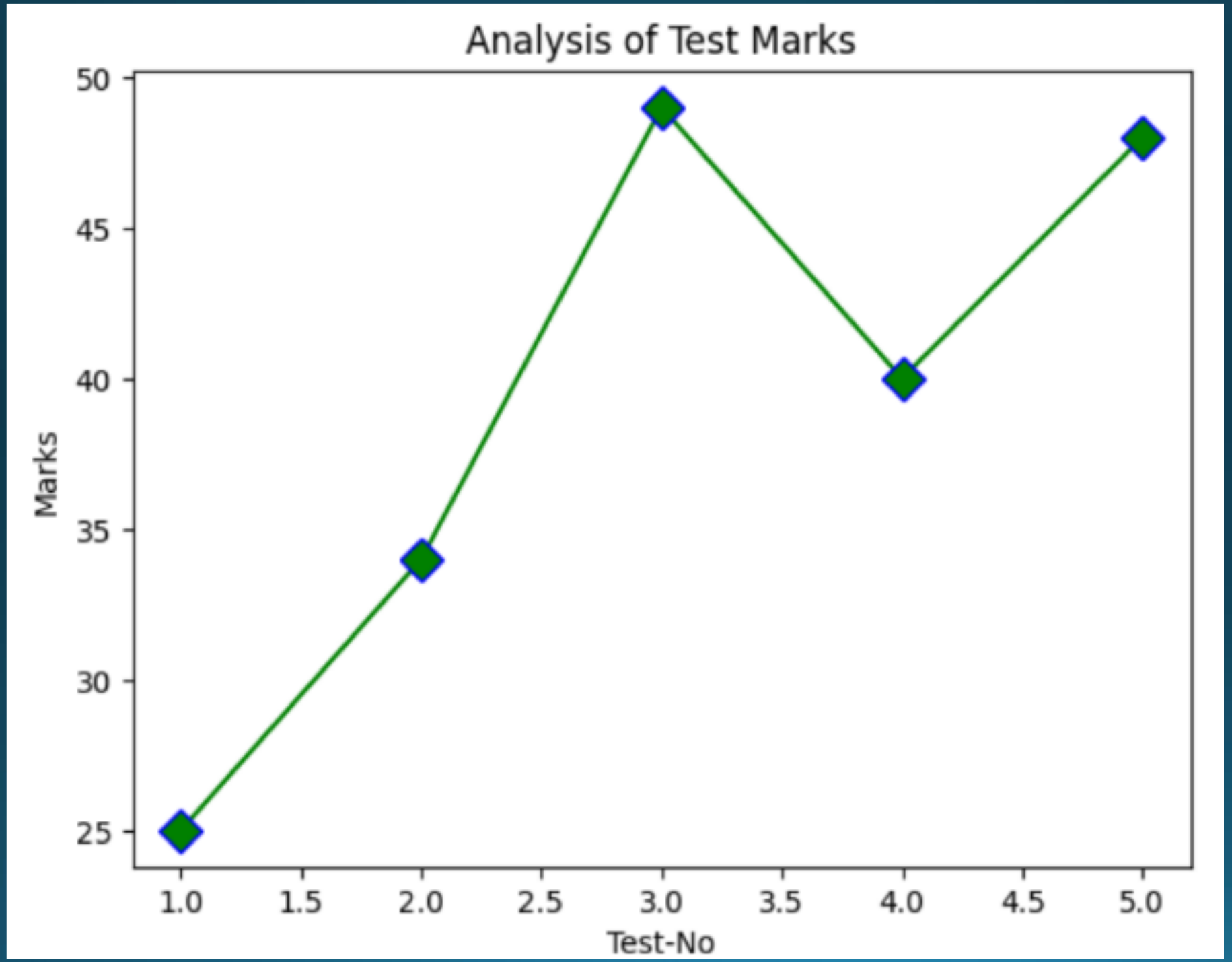
Line chart is plotted in python using the function plot ( ). Colour of the line can be mentioned by giving the colour codes inside the plot function.

Attributes of plot function which are used inside plot ( ) function are:

line width	Sets the width of the line
line style	determines the style of line (solid, dashed, dot, dashdot)
marker, markersize, markeredgecolor	determines the marker's shape, size and marker edge colour respectively

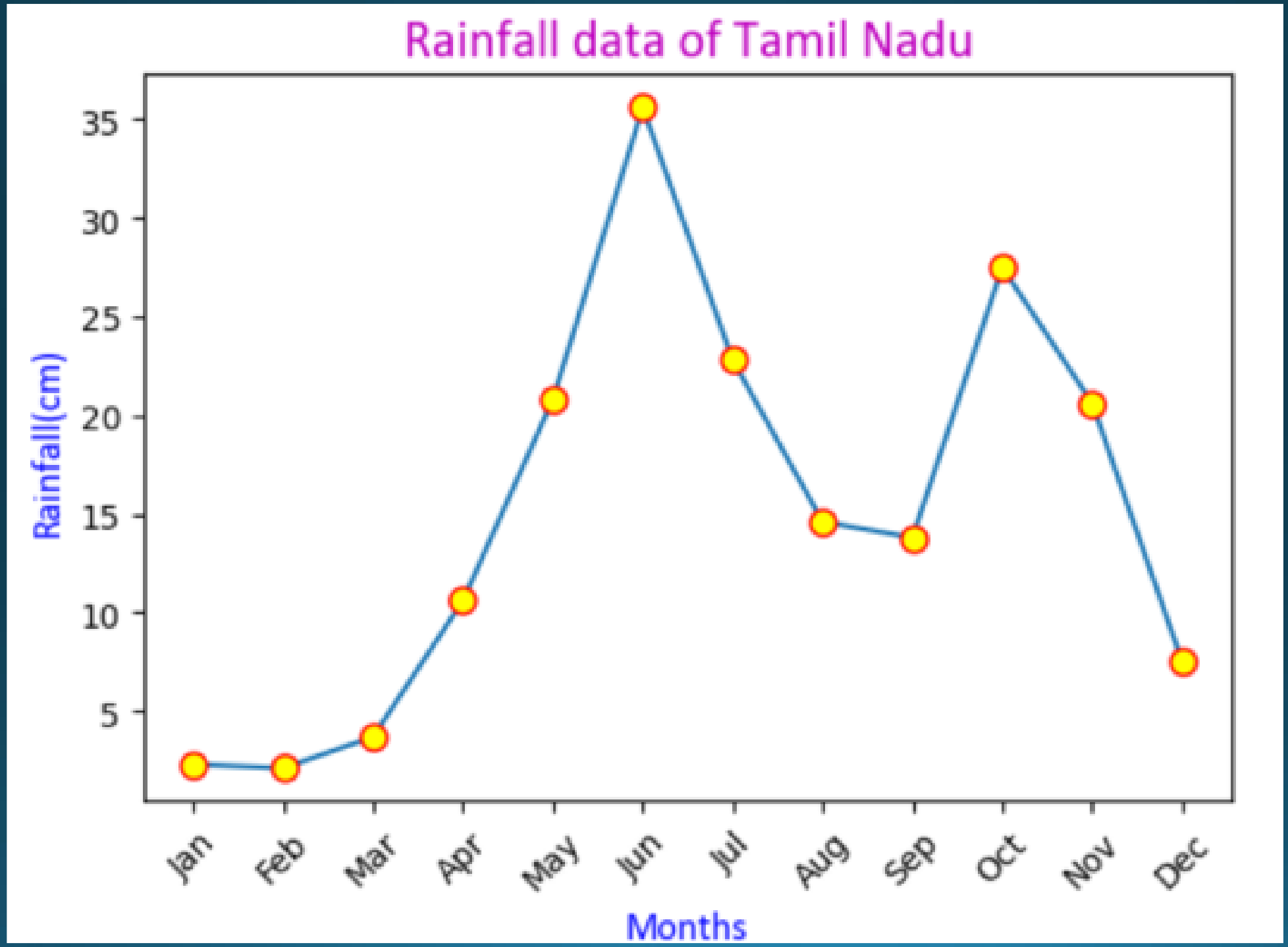
Program-5 Write a program to draw a line chart, we use plot function (use Example 1)

```
import matplotlib.pyplot as pl
Test=[1,2,3,4,5]
Marks=[25, 34, 49, 40, 48]

pl.title ("Analysis of Test Marks")
pl.xlabel("Test-No")
pl.ylabel("Marks")

pl.plot(Test, Marks,'g', marker='D', markersize=10, markeredgecolor='blue', linestyle='solid')
pl.show()
```


Analysis of Test Marks
50
45
40
35
30
25
Marks
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Test-No

Program -6 Write a program to draw a line chart to visualize the comparative rainfall data for 12 months in Tamil Nadu using the CSV file "rainfall.csv".

```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("rainfall.csv")
x=df['Months']
y=df['Rainfall(cm)']
plt.figure(figsize=(6,4))
plt.plot(x,y,marker='o', markersize=8, markeredgecolor='red', markerfacecolor='yellow')
plt.xticks(rotation = 45)
plt.xlabel("Months",fontname='Calibri',color='b',fontsize=12)
plt.ylabel("Rainfall(cm)",fontname='Calibri',color='b',fontsize=12)
plt.title("Rainfall data of Tamil Nadu",fontname='Calibri',color='m',fontsize=16)
plt.show()
```


Rainfall data of Tamil Nadu
Rainfall(cm)
35
30
25
20
15
10
5
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months

2. Bar Graph

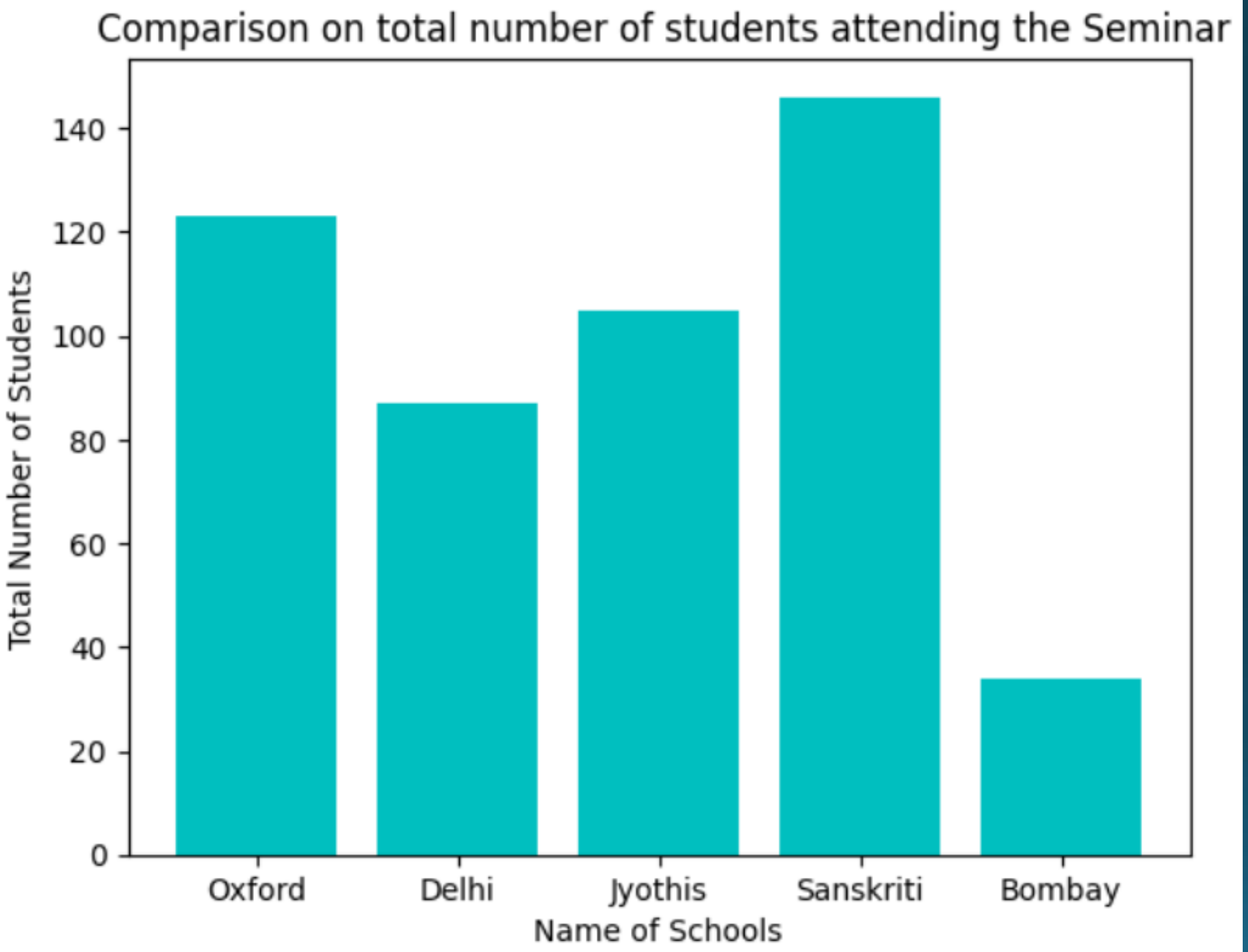
- A bar chart or bar graph is a graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

- It is a good way to show relative sizes, i.e., to show comparison between different categories.
-
- The relative sizes of the bars allow for easy comparison between different categories.

Example-6 Create a bar graph to illustrate the distribution of students from various schools who attended a seminar on “Deep Learning”. The total number of students from each school is provided below.

Oxford Public School	Delhi Public School	Jyothis Central School	Sanskriti School	Bombay Public School
123	87	105	146	34

```
import matplotlib.pyplot as pl
a=["Oxford", "Delhi", "Jyothis", "Sanskriti", "Bombay"]
b=[123, 87, 105, 146, 34]

pl.xlabel("Name of Schools")
pl.ylabel("Total Number of Students")

pl.title( " Comparison on total number of students attending the Seminar")
pl.bar(a,b,color='c')

pl.show()
```

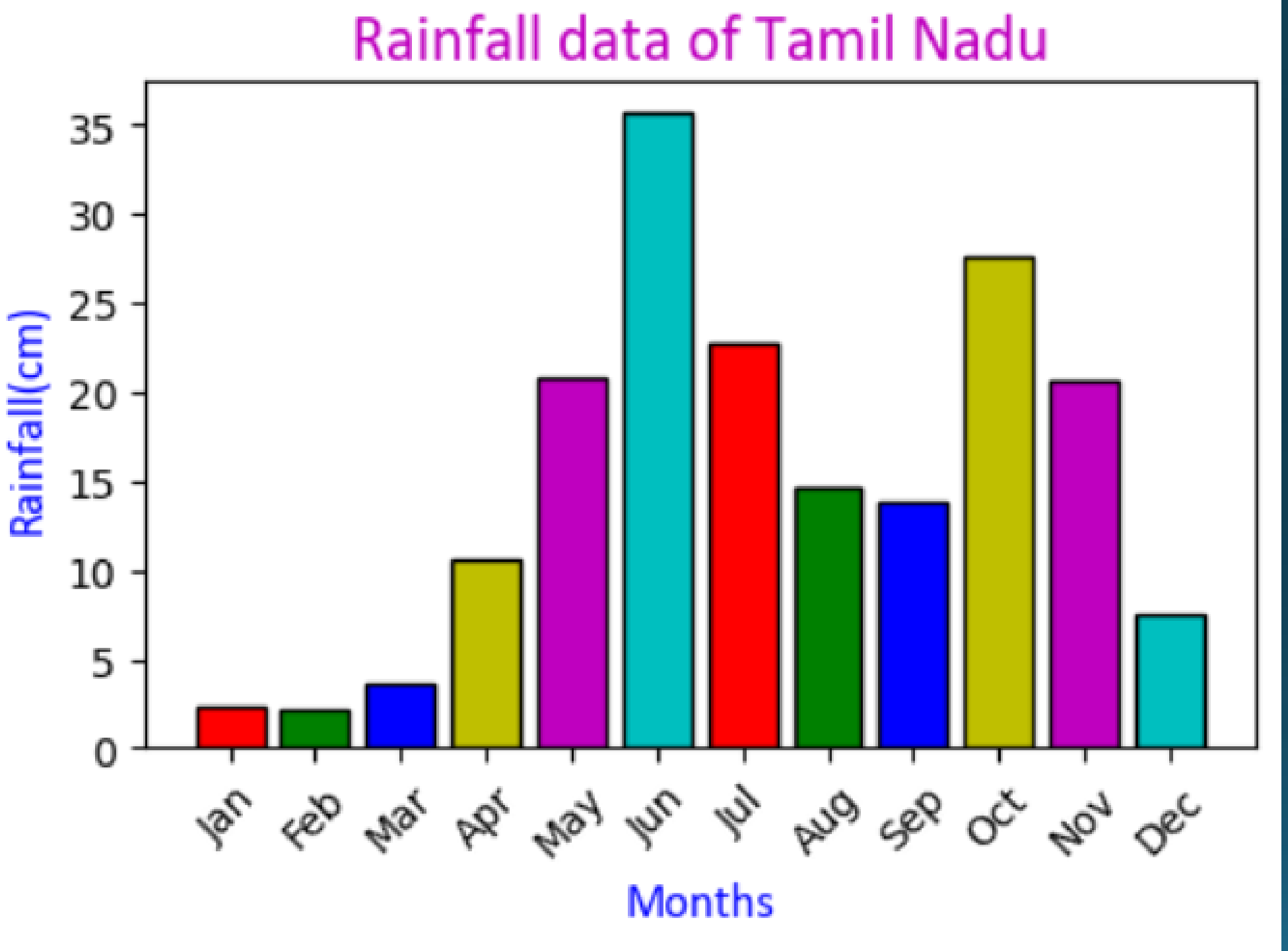

Comparison on total number of students attending the Seminar
Total Number of Students
0
20
40
60
80
100
120
140
Oxford
Delhi
Jyothis
Sanskriti
Bombay
Name of Schools

- Bar chart is plotted in python using the function bar ( ).
- Attributes of bar function which are used inside bar ( ) functions are:

color	determines the color of the bars
edgecolor	determines the colour of the bar edges
width	determines the width of the bars

Program – 7 Write a program to draw a bar chart to visualize the comparative rainfall data for 12 months in Tamil Nadu using the CSV file "rainfall.csv".

```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("rainfall.csv")
x=df['Months']
y=df['Rainfall(cm)']
plt.figure(figsize=(6,4))
c=['r','g','b','y','m','c']
plt.bar(x,y,color=c,edgecolor='k')
plt.xticks(rotation = 45)
plt.xlabel("Months",fontname='Calibri',color='b',fontsize=12)
plt.ylabel("Rainfall(cm)",fontname='Calibri',color='b',fontsize=12)
plt.title("Rainfall data of Tamil Nadu",fontname='Calibri',color='m',fontsize=16)
plt.show()
```


Rainfall data of Tamil Nadu
Rainfall(cm)
0
5
10
15
20
25
30
35
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months

3. Histogram

- Histograms are graphical representations of data distribution, with vertical rectangles depicting the frequencies of different value ranges.

- They are drawn on a natural scale, making it easy to interpret the central tendency, such as the mode, of the data.

- Despite their simplicity and ease of understanding, histograms have a limitation: they can only represent one data distribution per axis.

Solution:

To draw a histogram from this, we first need to organize the data into intervals. These intervals are also called logical ranges or bins. After computing the number of girls in each interval, draw the graph. Histogram is plotted in python using the function hist ( ).

```
import matplotlib.pyplot as pl
a=[141, 145, 142, 147, 144, 148, 141, 142, 149, 144, 143,
   149, 146, 141, 147, 142, 143]
pl.ylabel("Number of Girls")
pl.xlabel("Height")
pl.title( " Heights of Girls in class-XII")
pl.hist(a)
pl.show()
```


Heights of Girls in class-XII
Number of Girls
3.0
2.5
2.0
1.5
1.0
0.5
0.0
141
142
143
144
145
146
147
148
149
Height

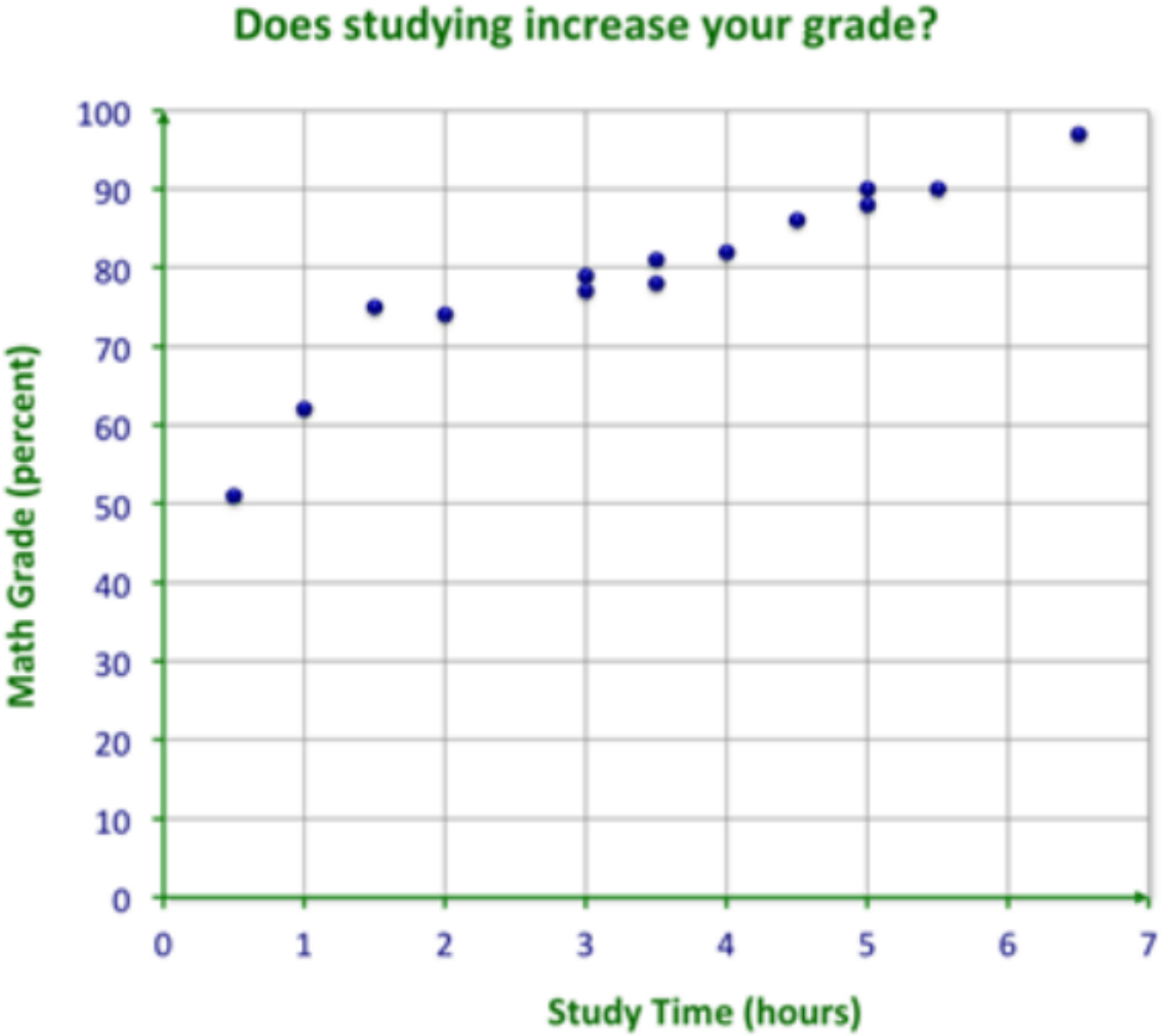
4. Scatter Graph

- Scatter plots visually represent relationships between two variables by plotting data points along both the x and y axes.

- They reveal correlations, whether positive or negative, within paired data, showcasing trends and patterns.

Example-8 A student had a hypothesis for a science project. He believed that the more the students studied Math, the better their math scores would be. He took a poll in which he asked students the average number of hours that they studied per week during a given semester. He then found out the overall percentage that they received in their Math classes. His data is shown in the table below

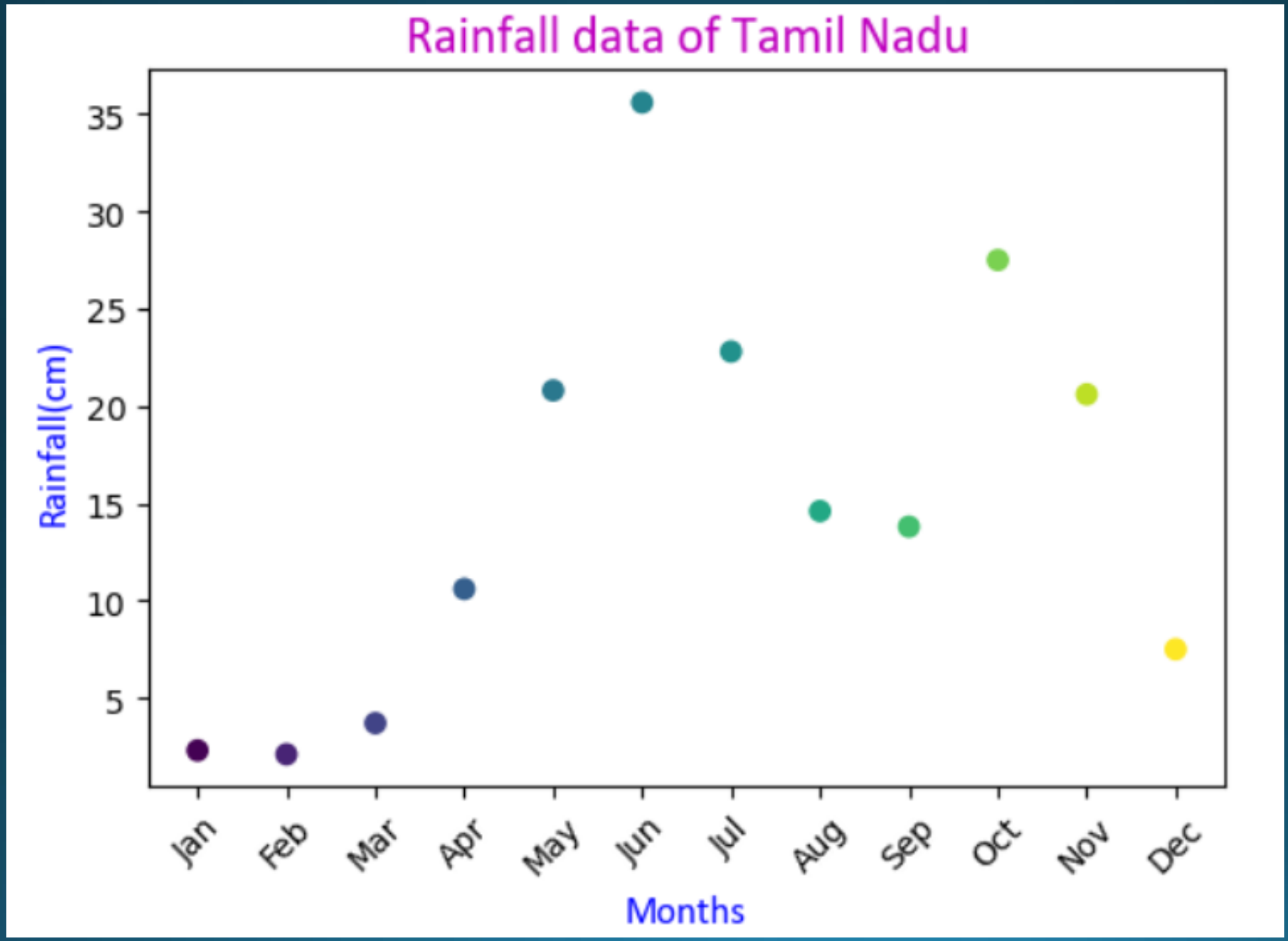
Study Time (Hours)	4	3.5	5	2	3	6.5	0.5	3.5	4.5	5
Maths Grade (%)	82	81	90	74	77	97	51	78	86	88


Does studying increase your grade?
Math Grade (percent)
100
90
80
70
60
50
40
30
20
10
0
0
1
2
3
4
5
6
7
Study Time (hours)

Scatterplot is plotted using the function scatter ( )

Program-8 Write a program to draw a scatter chart to visualize the comparative rainfall data for 12 months in Tamil Nadu using the CSV file "rainfall.csv".

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("rainfall.csv")
x=df['Months']
y=df['Rainfall(cm)']
plt.figure(figsize=(5,3))
plt.figure(figsize=(6,4))
colors = np.array([0, 10, 20, 30, 40, 45, 50,  60, 70, 80, 90, 100])
plt.scatter(x,y,c=colors,cmap='viridis')
plt.title("Rainfall data of Tamil Nadu",fontname='Calibri',color='m',fontsize=16)
plt.xticks(rotation = 45)
plt.xlabel("Months",fontname='Calibri',color='b',fontsize=12)
plt.ylabel("Rainfall(cm)",fontname='Calibri',color='b',fontsize=12)
plt.show()
```


Rainfall data of Tamil Nadu
Rainfall(cm)
35
30
25
20
15
10
5
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months

5. Pie Chart

- A pie chart is a circular graph divided into segments or sections, each representing a relative proportion or percentage of the total.

- Each segment resembles a slice of pie, hence the name.

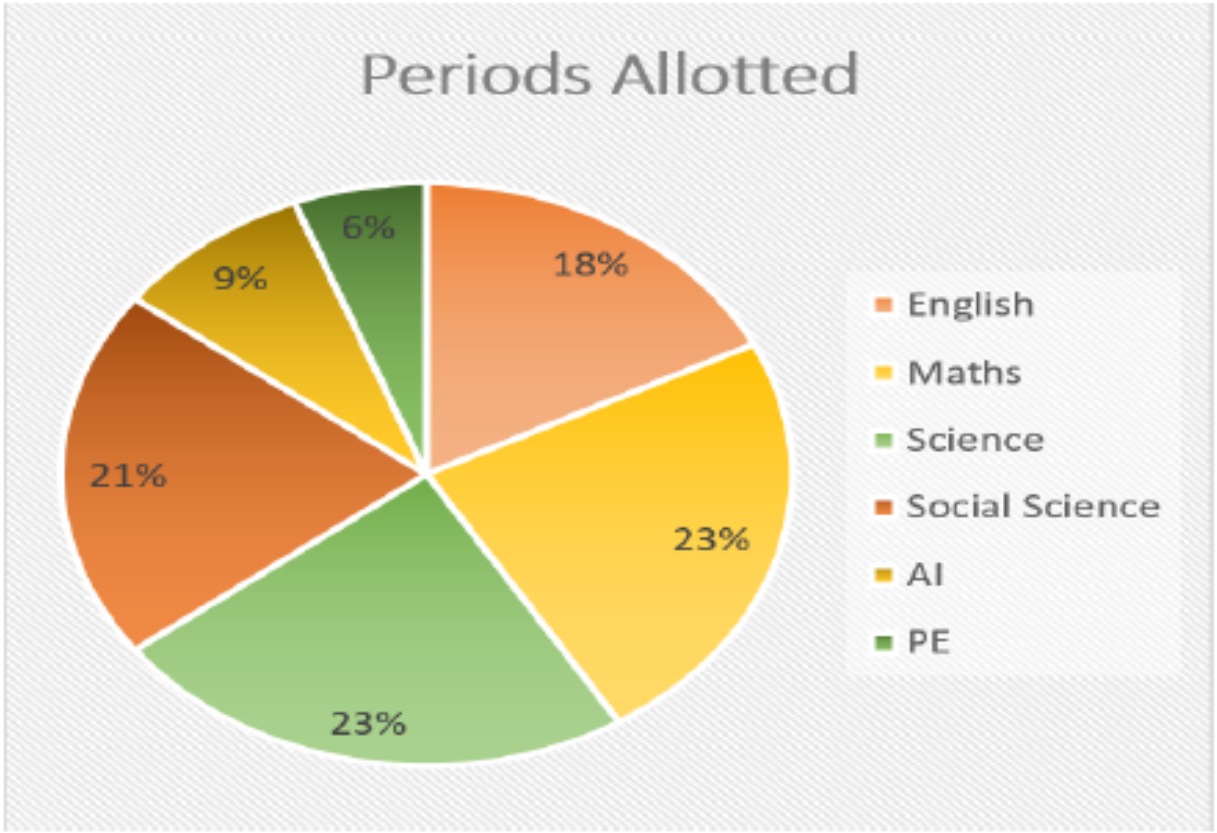
- Pie charts are commonly used to visualize data from a small table, but it is recommended to limit the number of categories to seven to maintain clarity.

- However, zero values cannot be depicted in pie charts.

Example-9

- Below given is a Pie chart drawn with the periods allotted for each subject in a week.

Subject	Periods Allotted
English	6
Maths	8
Science	8
Social Science	7
AI	3
PE	2



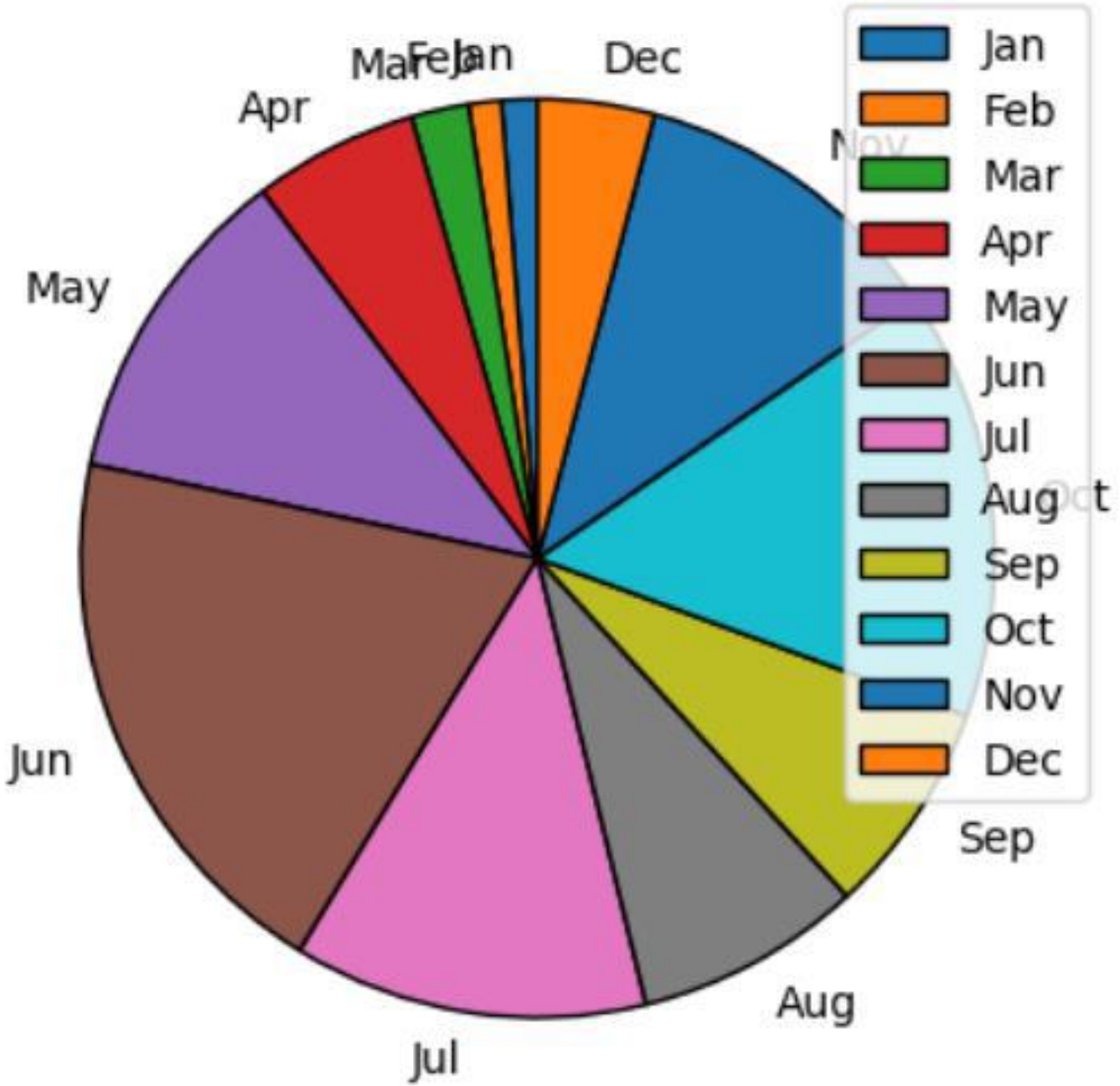
Pie Chart is plotted using the function pie ( )

Program-9

Write a program to draw a pie chart to visualize the comparative rainfall data for 12 months in Tamil Nadu using the CSV file "rainfall.csv".

```
import pandas as pd
import matplotlib.pyplot as plt
df=pd.read_csv("rainfall.csv")
x=df['Months']
y=df['Rainfall(cm)']
wp = { 'linewidth' : 1, 'edgecolor' : "black" }
plt.pie(y,labels=x, startangle=90,wedgeprops=wp)
plt. legend(loc='upper right')
plt.title("Rainfall data of Tamil Nadu",fontname='Calibri',color='m',fontsize=16)
plt.show()
```

Rainfall data of Tamil Nadu

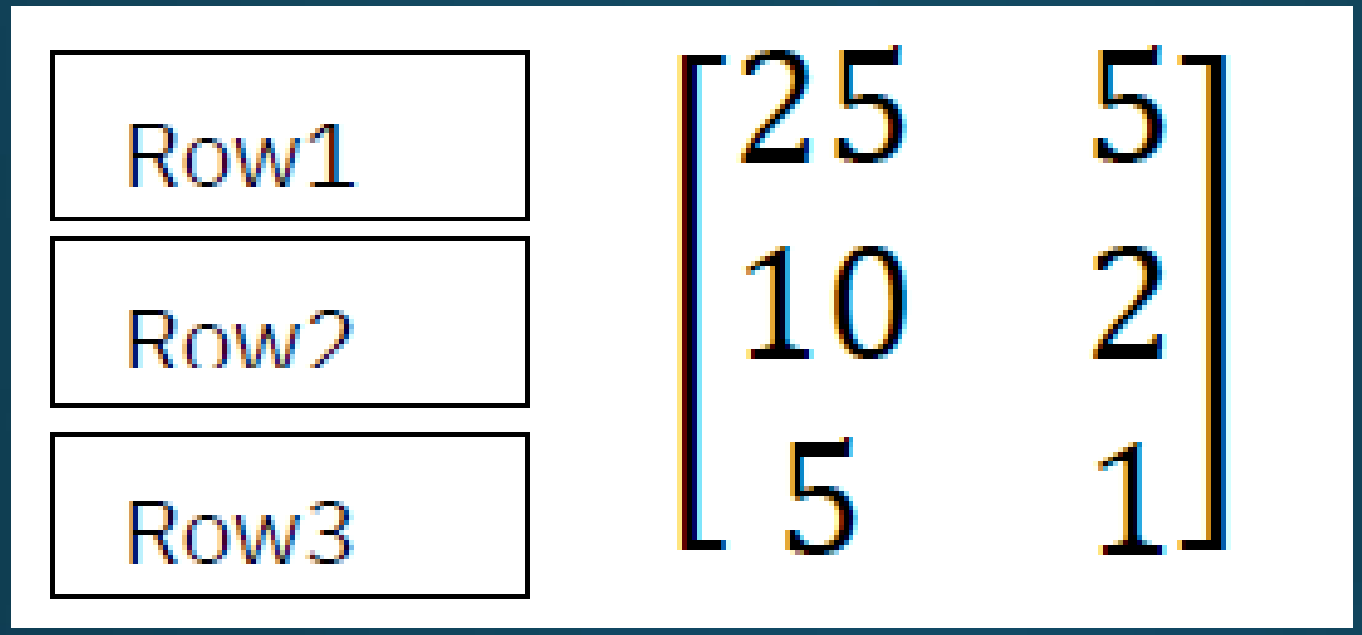


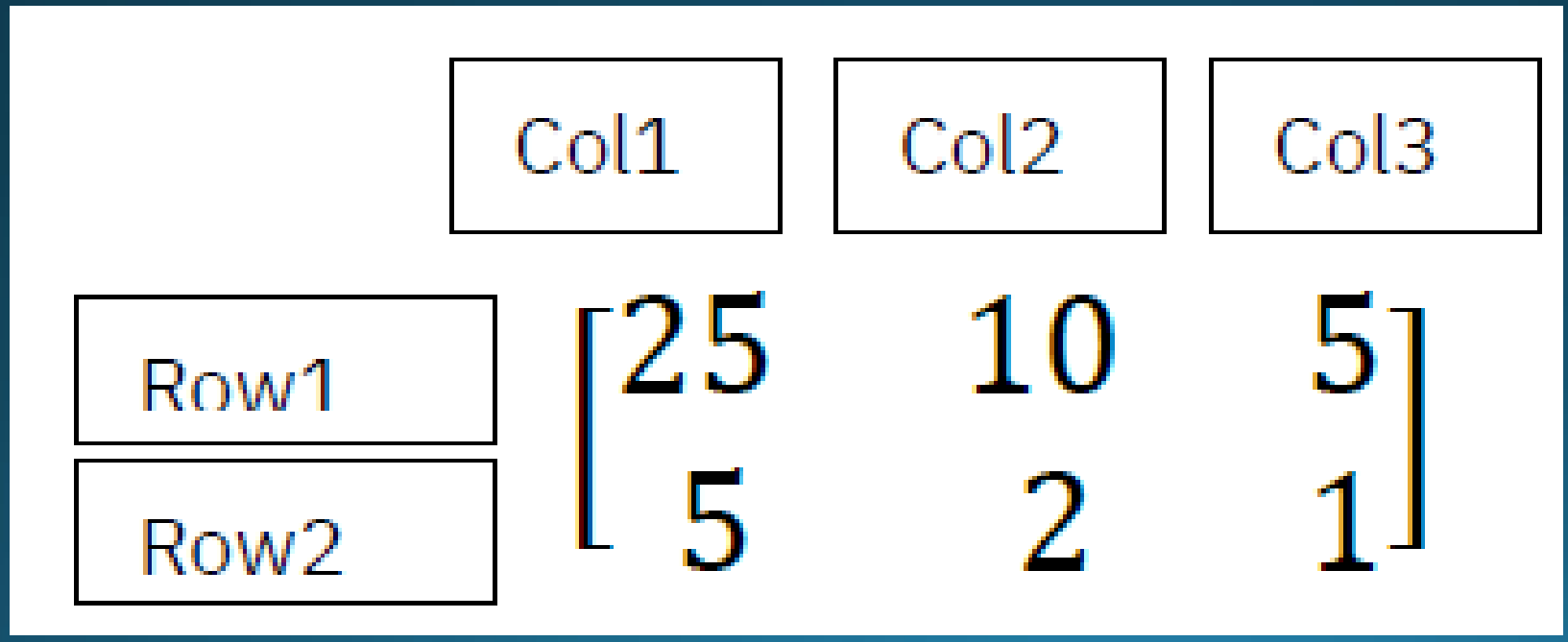
6. INTRODUCTION TO MATRICES

In mathematics, matrix (plural matrices) is a rectangular arrangement of numbers. The numbers are arranged in tabular form as rows and columns. Matrices play a huge role in computer vision domain of AI. On the computer, the image is represented as a combination of pixels. This is represented mathematically as matrices!

Let us understand with the help of an example:
Consider
Aditi bought 25 pencils 5 erasers
Adit bought 10 pencils 2 erasers
Manu bought 5 pencils 1 eraser

	Pencils	Erasers
Aditi	25	5
Adit	10	2
Manu	5	1


Row1
Row2
Row3
25 5
10 2
5 1


Col1
Col2
Col3
Row1
Row2
25 10 5
5 2 1

Order of a matrix

A matrix has m rows and n columns. It is called a matrix of order m × n or simply m × n matrix (read as an m by n matrix). So, the matrix A in the above example is a 3 × 2 matrix.

The number of elements are m x n => 3 x 2 = 6 elements. Each individual element is represented as aij where i represents row and j represents column. In general aij, is an element lying in the ith row and jth column. We can also call it as the (i, j)th element of the matrix.

$$P = \begin{bmatrix} a11 & a12 \\ a21 & a22 \\ a31 & a32 \end{bmatrix}$$

Operations on Matrices

1. Addition of matrices - the sum of two matrices is obtained by adding the corresponding elements of the given matrices. Also, the two matrices have to be of the same order. Example:

$$A = \begin{bmatrix} 3 & 2 \\ 4 & -1 \\ 2 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} 6 & 3 \\ 5 & 9 \\ 3 & 2 \end{bmatrix}$$

$$A+B = \begin{bmatrix} 3+6 & 2+3 \\ 4+5 & -1+9 \\ 2+3 & 0+2 \end{bmatrix} = \begin{bmatrix} 9 & 5 \\ 9 & 8 \\ 5 & 2 \end{bmatrix}$$

2. Difference of matrices - The difference A – B is defined as a matrix where each element is obtained by subtracting the corresponding elements (aij – bij). Matrices A and B must be of the same order. Example

$$A = \begin{bmatrix} -2 & 1 \\ 6 & 10 \\ 5 & 3 \end{bmatrix} \qquad B = \begin{bmatrix} -1 & 3 \\ 2 & 9 \\ 3 & 1 \end{bmatrix}$$

$$A\text{-}B = \begin{bmatrix} -2+1 & 1-3 \\ 6-2 & 10-9 \\ 5-3 & 3-1 \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ 4 & 1 \\ 2 & 2 \end{bmatrix}$$

3. Transpose of a matrix – a matrix obtained by interchanging the rows and columns. Transpose of a matrix A is denoted by A’ or AT. Example

$$A = \begin{bmatrix} 8 & 7 \\ 2 & 5 \\ 4 & 6 \end{bmatrix} \qquad A^T = \begin{bmatrix} 8 & 2 & 4 \\ 7 & 5 & 6 \end{bmatrix}$$

Order = 3x2 Order = 2x3

Applications of matrices in AI Matrices are used throughout the field of machine learning for computing:

•Image Processing – Digital images can be represented using matrices. Each pixel on the image has a numerical value. These values represent the intensity of the pixels. A grayscale or black-and-white image has pixel values ranging from 0 to 255. Smaller values closer to zero represent darker shades, whereas bigger ones closer to 255 represent lighter or white shades. So, in a computer, every image is kept as a matrix of integers called a Channel.

• Recommender systems use matrices to relate between users and the purchased or viewed product(s)

• In Natural Language processing, vectors depict the distribution of a particular word in a document. Vectors are one-dimensional matrices.

7. DATA PREPROCESSING-

• Imagine you have a giant bag of mixed candies. How can we sort through the candies to make it easier to find the Flavors we want? (This relates to data cleaning and organization)

• Sometimes data can be like a bag of candy with a few weird pieces mixed in. How can we make sure all the data makes sense of what we're trying to learn from it? (This relates to identifying and handling outliers and inconsistencies)

- Data preprocessing is a crucial step in the machine learning process aimed at making datasets more machine learning-friendly.

- It involves several processes to clean, transform, reduce, integrate, and normalize data:

1. Data Cleaning

1.Missing Data:

- Missing data occurs when values are absent from the dataset, which can happen due to various reasons. Strategies for handling missing data include deleting rows or columns with missing values, inputting missing values with estimates, or using algorithms that can handle missing data.

2.Outliers:

- Outliers are data points that significantly differ from the rest of the data, often due to errors or rare events. Dealing with outliers involves identifying and removing them, transforming the data, or using robust statistical methods to reduce their impact.

3.Inconsistent Data

- Data with typographical errors, different data types etc. are corrected and consistency among the data is observed.

4.Duplicate Data

- Duplicate data will be identified and removed to ensure data integrity.

2. Data Transformation

Categorical variables are converted to Numerical variable. New features are identified and existing features are modified if needed

3. Data Reduction

Dimensionality reduction, i.e. reducing the number of features of data set is done. If data set is too large to handle sampling techniques are applied

4. Data Integration and Normalization

If data is stored in multiple sources or formats, they are merged or aggregated together. Then the data is normalized to ensure that all features have a similar scale and distribution which can improve machine learning models.

5. Feature Selection

The most relevant features that contribute the most to the target variable are selected and irrelevant data are removed.

8. DATA IN MODELLING & EVALUATION

• Imagine you're lost in a new city. You have a map, but it might be helpful to have a separate practice map to try figuring things out before using it on the actual streets. Why do you think splitting data into training and testing sets might be similar to this idea? (This question connects the concept of data splitting to a familiar scenario and highlights the purpose of each set)

• When making a decision, do you consider all the information available or just some of it? Why do you think using techniques like cross-validation might be important when evaluating how well a machine learning model performs? (This question relates the importance of evaluating models to make informed decisions and introduces the concept of cross-validation)

- After the data is pre-processed, it is splitted into two -- Training data set and Testing data set.

- The training set is used to train the machine learning models, while the testing set is used to evaluate the performance of the trained models.

- While modelling, appropriate machine learning algorithms are chosen based on the nature of the problem (e.g., classification, regression, clustering) and the characteristics of the dataset.

- Techniques such as train-test split, cross-validation, and error analysis are employed to estimate the model's generalization ability and identify areas for improvement.

- Train-Test Split trains the model with its training set and evaluates using the test set. Cross Validation ensures that the model's performance is consistent across different subsets of the data.

- Different types of evaluation techniques are applied on the model depending on the data. For classification problems, metrics like accuracy, precision, recall, F1-score, and ROC curve are commonly used.
- For regression problems, metrics like mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared are often used.

- In today's world, knowing how to work with data is important. As artificial intelligence becomes more and more common, understanding data helps us use information better.

- It is like having a map to find your way through a big city. Being good with data helps us make smart decisions and use technology wisely.

THANK YOU