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.

EXERCISES:

I. Multiple-choice questions

1. Which of the following best defines data literacy?
   A) The ability to read and write data
   B) The ability to find and use data effectively
   C) The ability to analyse data using AI
   D) The ability to collect and store data securely

2. What is the purpose of data preprocessing?
   A) To make data more complex
   B) To make data less accessible
   C) To clean and prepare data for analysis
   D) To increase the size of the dataset

3. How can missing data be handled in a dataset?
   A) By ignoring it
   B) By replacing missing values with estimates
   C) By deleting rows or columns with missing values
   D) By converting missing values to zero

4. Which of the following statements about the quantity of data needed for machine learning projects is true?
   A) More data is always better for good predictions.

B) Small batches of data are sufficient for complex models.
C) Data quantity depends solely on the number of features.
D) Data diversity is not essential for model performance.

5. Which of the following is an example of a primary source of data collection?
A) Web scraping B) Social media data tracking
C) Surveys D) Kaggle datasets

6. What method of data collection involves direct communication with individuals or groups to gather information?
A) Observations B) Experiments C) Interviews D) Marketing campaigns

7. Which of the following is an example of ratio scale data?
A) Grading students' exam papers as "A," "B," "C," "D," and "F"
B) Measuring the temperature in Celsius
C) Rating a meal at a restaurant as "unpalatable," "unappetizing," "just okay," "tasty," and "delicious"
D) Recording the weight of a person in kilograms

8. What is the distinguishing feature of ratio scale data?
A) It involves categories without a specific order
B) It has a zero point and allows for ratios to be calculated
C) It involves categories with a strict order but no measurable differences between categories
D) It has a definite order, but the differences between categories cannot be measured

9. Which statistical measure is most suitable for data sets with evenly spread values and no exceptionally high or low values?
A) Mean B) Median C) Mode D) Variance

10. What is the term used to describe the graphical or pictorial representation of data?
A) Statistical summary B) Data organization
C) Data visualization D) Data interpretation

ANSWERS
1. B) The ability to find and use data effectively
2. C) To clean and prepare data for analysis
3. B) By replacing missing values with estimates
4. A) More data is always better for good predictions.
5. C) Surveys
6. C) Interviews
7. C) Rating a meal at a restaurant as "unpalatable," "unappetizing," "just okay," "tasty," and "delicious"
8. B) It has a zero point and allows for ratios to be calculated
9. A) Mean
10. C) Data visualization

II. Short answer questions:

1. Explain the concept of data literacy and its importance in today's digital age.

Data literacy refers to the ability to find, understand, and use data effectively. In today's digital age, where data is abundant, data literacy is crucial for making informed decisions, understanding trends, and solving complex problems. It helps individuals and organizations extract valuable insights from data to drive innovation and growth.

2. What is data preprocessing?

Data preprocessing is the process of cleaning and preparing raw data before it is used for analysis or modelling. It involves handling missing data and outliers to ensure the data is accurate and reliable.

3.What is data visualization and why is it important?

Data visualization is the graphical representation of data to help people understand the significance of data by summarizing and presenting it in a visual form such as charts, graphs, or maps.

Data visualization is important because it allows for the exploration and understanding of data patterns, trends, and outliers that may not be apparent in raw data. It helps in making data-driven decisions and communicating information clearly and efficiently.

4. How does a line graph differ from a bar graph?

A line graph is used to show trends over time with continuous data, while a bar graph is used to compare different categories of data with discrete values.

5. When would you use a scatter plot?

A scatter plot is used to show the relationship between two variables in a set of paired data, helping to identify correlations or trends between the variables.

6. What is data?

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.

7. What do you mean by web scraping?

Web scraping is the process of using bots to extract content and data from a website. Web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.

8. If a matrix has 6 elements, what are the possible orders it can have?

Answer: 4 orders – (1x6), (6x1), (2x3) and (3x2)

9.Construct a 3x2 matrix where each element is given by $a_{ij} = i * j$

Answer: 3x2 matrix means 6 elements

a11 = 1x1	a12 = 1x2
a21=2x1	a22=2x2
a31=3x1	a32=3x2

Putting all elements in matrix form we get:

$$A=\begin{bmatrix}1 & 2\\ 2 & 4\\ 3 & 6\end{bmatrix}$$

9. Find the transpose of the matrix B = $\begin{bmatrix}5 & -1 & 4\\ 2 & 3 & 6\end{bmatrix}$

Answer: B = $\begin{bmatrix}5 & 2\\ -1 & 3\\ 4 & 6\end{bmatrix}$

III. Long answer questions:

1. Discuss the advantages and limitations of using a pie chart in data visualization. Provide examples to illustrate your points.

One advantage of using a pie chart is that it can effectively show the proportion of each category in a dataset. For example, a pie chart can be used to visualize the market share of different companies in a specific industry. However, pie charts have limitations, such as difficulty in comparing multiple datasets or showing trends over time. For example, a pie chart would not be suitable for visualizing changes in sales over the course of a year, as it cannot effectively convey this type of information.

2. Explain the terms mean, median and mode.

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.

3. Explain the four levels of measurement.

Four levels of measurements are Nominal, Ordinal, Interval and Ratio

The nominal level of measurement is the simplest or lowest of the four ways to characterize data. Nominal means "in name only". Colours of eyes, yes or no responses to a survey, gender, smartphone companies, etc all deal with the nominal level of measurement.

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. The options is Ordinal and the data is qualitative,. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the interval scale is similar to ordinal level data because it has a definite ordering but in interval scale data does not have a starting point i.e. zero value. Temperature scales like Celsius (oC) and Fahrenheit (F) are measured by using the interval scale. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. We can add, subtract, divide and multiply the two ratio level variables. Eg: Weight of a person. It has a real zero point. Hence it can be considered as ratio value.

4. Given the matrices A and B. Calculate A+ B and B – A.

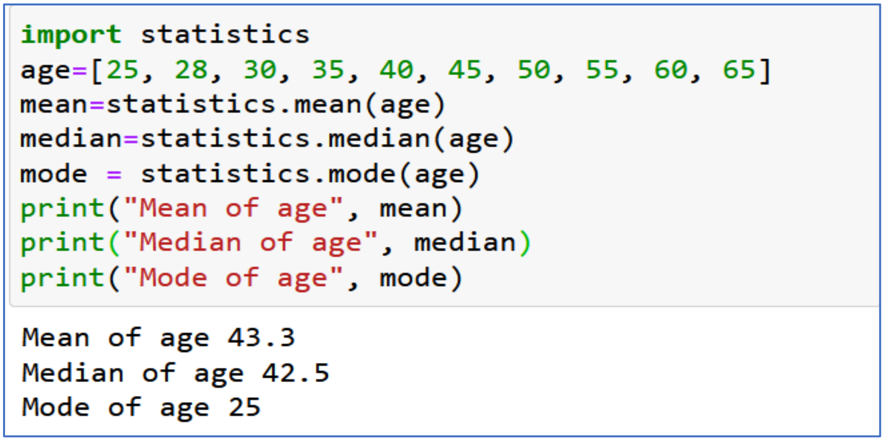
$$A = \begin{bmatrix} 2 & 7 \\ 4 & 12 \\ 15 & -3 \end{bmatrix} \qquad B = \begin{bmatrix} -2 & 4 \\ 12 & 6 \\ 7 & 0 \end{bmatrix}$$

Answer:

$$A+B = \begin{bmatrix} 0 & 11 \\ 16 & 18 \\ 22 & -3 \end{bmatrix} \qquad B - A = \begin{bmatrix} -4 & -3 \\ 8 & -6 \\ -8 & 3 \end{bmatrix}$$

IV. Python Programs

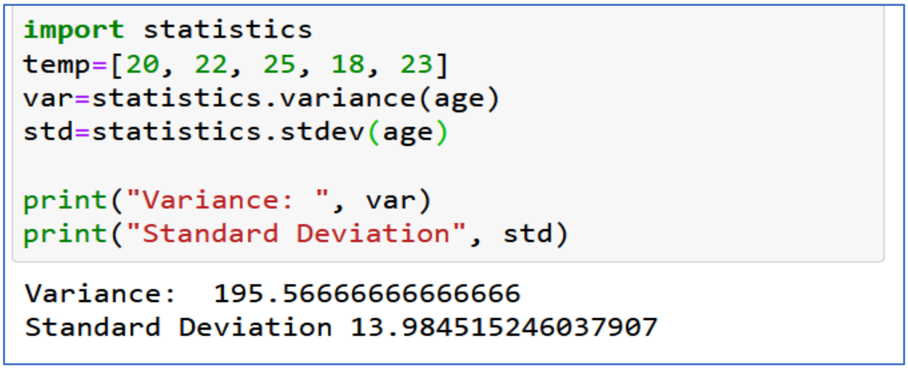
1. The ages of a group of people in a community are: 25, 28, 30, 35, 40, 45, 50, 55, 60, 65. Write a program to calculate the mean, median, and mode of the ages.



```
import statistics
age=[25, 28, 30, 35, 40, 45, 50, 55, 60, 65]
mean=statistics.mean(age)
median=statistics.median(age)
mode = statistics.mode(age)
print("Mean of age", mean)
print("Median of age", median)
print("Mode of age", mode)
```

```
Mean of age 43.3
Median of age 42.5
Mode of age 25
```

2. A company recorded the daily temperatures (in degrees Celsius) for five consecutive days: 20°C, 22°C, 25°C, 18°C, and 23°C. Determine the variance and standard deviation of the temperatures.

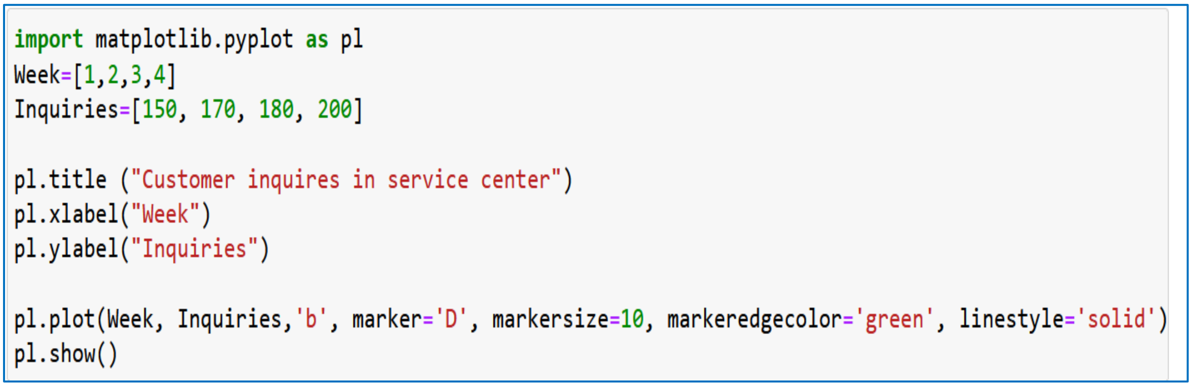


```
import statistics
temp=[20, 22, 25, 18, 23]
var=statistics.variance(age)
std=statistics.stdev(age)

print("Variance: ", var)
print("Standard Deviation", std)
```

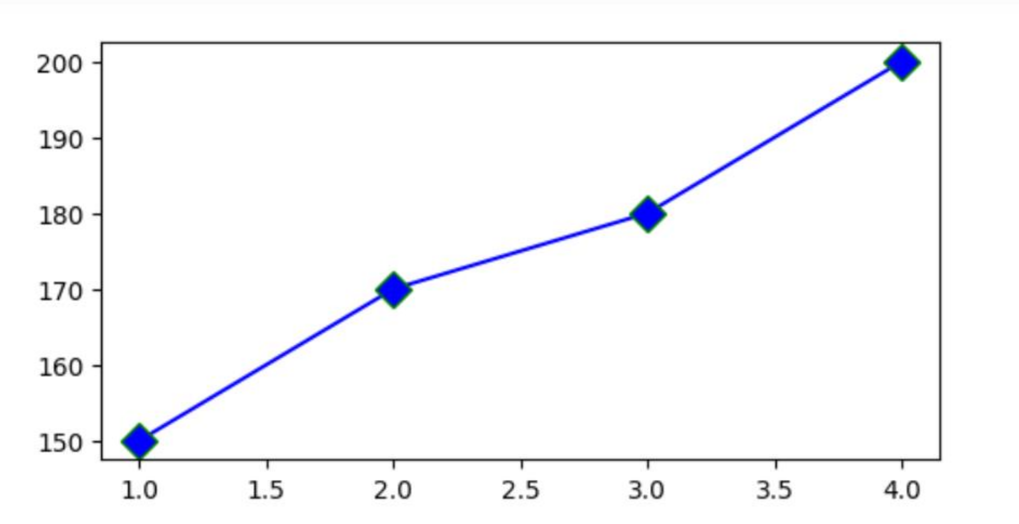
```
Variance:  195.56666666666666
Standard Deviation 13.984515246037907
```

3. Plot a line chart representing the weekly number of customer inquiries received by a customer service center:

- Week 1: 150 inquiries
- Week 2: 170 inquiries
- Week 3: 180 inquiries
- Week 4: 200 inquiries

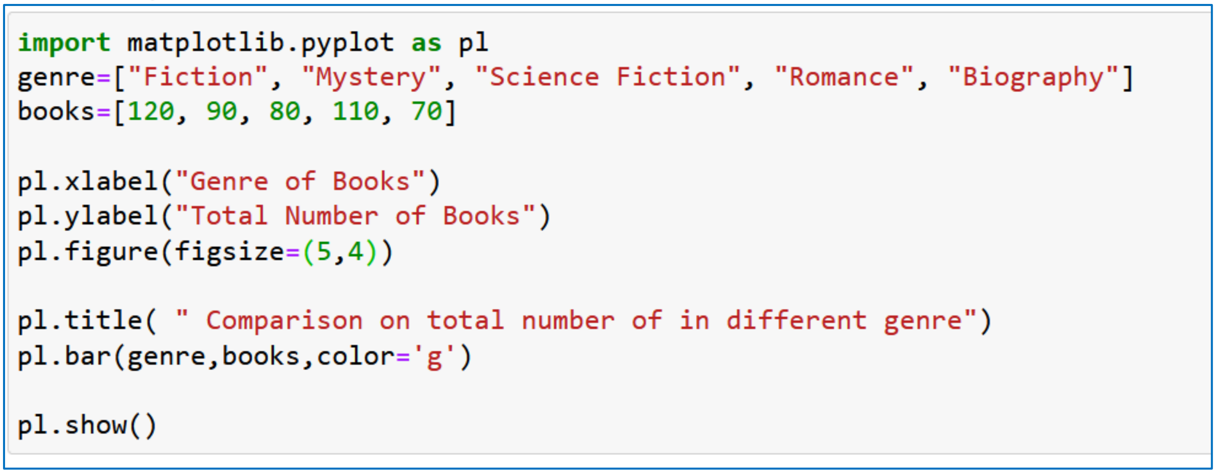


```
import matplotlib.pyplot as pl
Week=[1,2,3,4]
Inquiries=[150, 170, 180, 200]

pl.title ("Customer inquires in service center")
pl.xlabel("Week")
pl.ylabel("Inquiries")

pl.plot(Week, Inquiries,'b', marker='D', markersize=10, markeredgecolor='green', linestyle='solid')
pl.show()
```

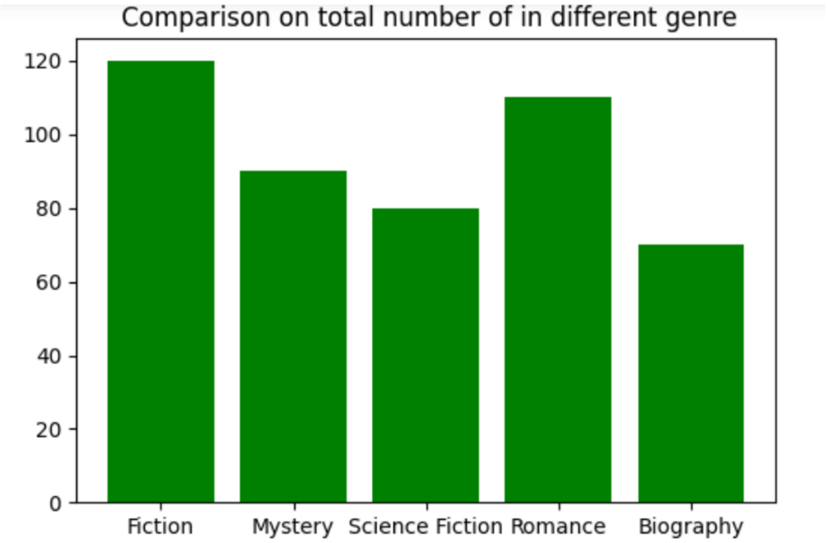


4. Plot a bar chart representing the number of books sold by different genres in a bookstore:

- Fiction: 120 books
- Mystery: 90 books
- Science Fiction: 80 books
- Romance: 110 books
- Biography: 70 books

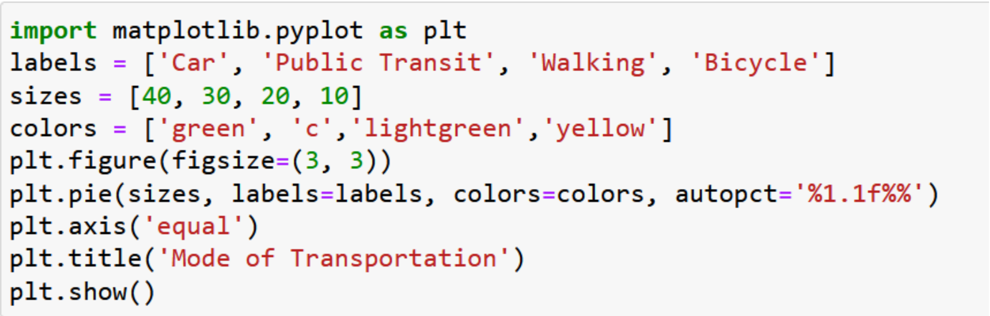


```python
import matplotlib.pyplot as pl
genre=["Fiction", "Mystery", "Science Fiction", "Romance", "Biography"]
books=[120, 90, 80, 110, 70]

pl.xlabel("Genre of Books")
pl.ylabel("Total Number of Books")
pl.figure(figsize=(5,4))

pl.title( " Comparison on total number of in different genre")
pl.bar(genre,books,color='g')

pl.show()
```

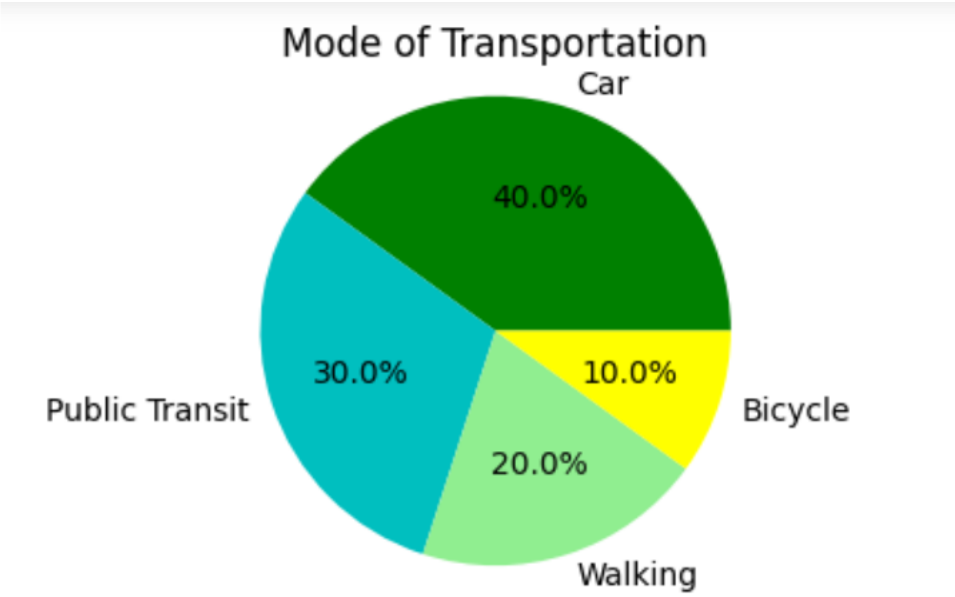


5. Visualize the distribution of different types of transportation used by commuters in a city using a pie chart:

- Car: 40%
- Public Transit: 30%
- Walking: 20%
- Bicycle: 10%



```python
import matplotlib.pyplot as plt
labels = ['Car', 'Public Transit', 'Walking', 'Bicycle']
sizes = [40, 30, 20, 10]
colors = ['green', 'c','lightgreen','yellow']
plt.figure(figsize=(3, 3))
plt.pie(sizes, labels=labels, colors=colors, autopct='%1.1f%%')
plt.axis('equal')
plt.title('Mode of Transportation')
plt.show()
```



V. Competency Based Questions

1. Anakha is working on a project involving mood analysis of individuals experiencing depression. She plans to visit hospitals, yoga instructors, mental health care providers, and healthcare institutes. How can Anakha collect data from these organizations and individuals for her project?
   Ans: Surveys, Interviews, Questionnaire

2. Kalaimathi and her friend are planning to build an admission chatbot for her school. They have collected data from different sources with the help of their friends. Now, they need to arrange it in proper order to analyse it. They plan to classify the data based on the levels of measurement. The data they have collected is as follows: *Student Name, Age, Gender, GPA (Grade Point Average), ECA (Extra-Curricular activities), Place, Parent Name, Parent Education level, Distance from school, Fees, Interview rating, No. of years in last school, Admission Test score*
   Ans:

NOMINAL	**ORDINAL**	**INTERVAL**	**RATIO**
Student Name	Parent Educational level	No. of years in last school	Age
Gender	Interview rating	Admission test score	GPA
Place		Distance from school to home	ECA
Parent Name			Fees

3. During a sales analysis, metrics such as sales revenue per month, average sales revenue, and the most popular products sold are examined to comprehend overall performance and aid in decision-making regarding marketing strategies, inventory management, and resource allocation. Which measurements in statistics will facilitate the sales analysis process?
   Ans: Mean, Median & Mode

4. Selvan's Textiles operates multiple showrooms in a city. Fahad is working on a project to predict the annual sales percentage for the upcoming year. He intends to analyse the trend of sales percentage over time and also the sales percentages in different regions. To facilitate data analysis, he plans to visualize the data using graphs.
    a. Which type of graph would be most appropriate for visualizing the trend of sales percentage over time?
    b. Which type of graph would be most suitable for comparing sales percentages across different regions?

   Ans:
    a. Line Chart
    b. Pie Chart

5. Akshith wrote a program to visualize the data analysis of five test and marks got.

```
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')
```

The program did not have any errors. But the line graph was not showing up. Could you find the reason why the graph is not shown even though the program has no errors?

Ans:

In order to view the graph we have to use show( ) function in the program. If we write pl.show( ), the graph window will be displayed.