

Student Learning Outcomes

By the end of this chapter, you will be able to:

- Explain the concept and importance of data science in everyday life.
- Describe each step in the Data Science Life Cycle (DSLCL).
- Identify and define real-world problems that can be solved using data.
- Collect data using surveys, online forms, and observational techniques.
- Perform data cleaning and validation to improve accuracy.
- Use simple SQL queries (SELECT, INSERT) to retrieve and store data from databases.
- Analyze data using basic algorithms (sorting, classification, clustering).
- Create bar charts and dashboards to visually represent data insights.
- Communicate results clearly using charts and written summaries.

Subjective Type Questions

Long Questions prepared in the light of new Examination Techniques
(Knowledge, Understanding, Application, Analytical & Conceptual)

Introduction

Q. Why Data science is important now a days?

Ans: Data Science:

Data science is the process of using data to understand problems, discover patterns, and make smart decisions. In the current digital era, data is everywhere-i.e., from surveys to mobile apps, social media, online shopping, and health services.

Importance:

Data Science helps us turn this raw data into useful information. Whether it is predicting weather, understanding interests of students, or improving healthcare, Data Science is transforming the way we live, learn, and work. Organizations across the whole world use data to improve decision-making, save costs, and create better services.

This chapter introduces the Data Science Life Cycle (DSLCL), which includes understanding the problem, collecting and cleaning data, analysing and interpreting data, and finally communicating the results using charts and visuals. Each step builds critical thinking and problem-solving skills.

Students will also learn how to collect, clean, analyse, interpret and visualize data using simple tools like surveys, spreadsheets, and charts. The chapter includes real-world examples, and hands-on activities that help them apply Data Science in classroom and daily life.

Q.1. Define Data Science? Also describe its importance?

Ans:

Data science is the process of using data to understand problems, discover patterns, and make smart decisions. Whether we are using a mobile phone, searching on Google, shopping online, or even walking with a fitness tracker, data is being collected.

Purpose of Data Science:

Data Science is the field where we use data to find patterns, answer questions, and help people or businesses make better decisions. It involves collecting, organizing, analyzing, and visualizing data.

Importance of data in decision making:

Data plays a big role in helping people and organizations make better decisions. Without data, decisions may be based on guesses or personal opinions which can lead to mistakes.

Example:

if the school administration desires to start a new sports club, they can first collect data on which sports students enjoy most. This helps them make a smart and fair choice.

Q.2. Explain real life applications of data science.

Ans: Real Life Applications of Data Science:

Social media: Applications like Facebook, YouTube and Instagram use data to show us posts and videos we might like.

Online shopping: E-commerce websites suggest products using data of our previous purchases.

Healthcare: Hospitals use data science to predict diseases and manage patient care more effectively.

Transport: Applications like Google Maps use traffic data to suggest the fastest routes.

Education: In schools, data from student attendance and test results is used to improve learning strategies and track progress.

Q.3. Define DSLC and write the name of its steps?

Ans:

Data Science is not covered in a single step. It follows a series of steps that are helpful in solving a problem. This is what we call Data Science Life Cycle (DSLCL). DSLC usually comprises of the following steps:

- Understanding the problem
- Collecting Data
- Cleaning and Validating data

- Analyzing Data
- Interpreting Data
- Visualizing and Communication Results

Q.4. Explain Understanding the problem in DSLC?

Ans:

Understanding the problem is the most important step in DSLC. Before we start talking about collecting data, it is necessary to clearly understand the problem under consideration. In case, we do not understand the problem properly, we may collect the wrong data resulting in useless or wrong results. In this initial step, we need to ask questions such as:

- What are we trying to find out?
- Why is this problem important?
- Who will use the results?

Sub-sections of Understanding the Problem:

Identifying and Defining a Problem:

First, we need to identify a problem that we want to solve using data. The problem can be related where decisions are made based on information. Once a problem is identified, then we must define it clearly and simply. When defining a problem, we need to consider the following:

- Use clear and simple words
- Focus on one specific question
- Avoid unclear or general terms.

TIDBIT

A good problem is one that can be answered by collecting and analyzing data.
Setting Clear objective:

Once we have identified and defined the problem, we must decide what we want to achieve. These are called our objectives which help in guiding the rest of the data science process. They state:

- What kind of data we need
- What questions we will ask
- What results we are looking for.

TIDBIT

Write your objectives like a checklist. This helps keep your project focused.

Examples: School, Health, Business related Problems and Objectives

This subsection covers a few realistic examples to understand the concept in a better way. Table 5.1 provides example problems and objectives related to different domains such as school, health and business.

Table 5.1: Example Problems and Objectives

Domain	Example Problem	Objectives
School	Many students arrive late at school	Find out how many students arrive late in month.
		Identify reasons for being late (e.g., transport, distance, emergency, weather issues).
		Suggest solutions to reduce late arrivals.
Health	Many students are not drinking enough water during school hours.	Collect data on how much water students drink daily.
		Find out reasons for low water intake.
Business	A school canteen wants to know which snacks students like the most.	Recommend steps to promote healthy habits.
		Survey students on their favorite snacks.
		Compare popularity and sales data. Help the canteen manager make smart ordering decisions.

REMEMBER:

1. The first step in any data science project is to understand the problem.
2. We must define the problem clearly and set specific goals.
3. Good problems are solvable using data.
4. Clear objectives help in collecting the right data and reaching useful results.

Q.5. Define data collection and also explain the methods of data collection?

Ans:

Data is like raw material, and we must gather the right kind of data before we can clean, store, or analyze it. For collecting the data, we use the following steps:

- Data Collection methods
- Choosing the right source
- Assessing data quality and bias

Data Collection Methods:

There are many ways to collect data, depending on the problem and available tools.

Surveys (Online/paper based):

Surveys are the easiest and most common method. We ask people questions and record their answers.

Example:
To find out which sports students enjoy, we can create a survey with questions using Google form and share its link with students for response.

DO YOU KNOW

Surveys can be done using Google Forms (online), and Printed paper forms (manual).

APIs and Online Sources:

Application Programmer Interface (APIs) allow software collect data from websites or applications (apps in short) automatically.

Example:

A school may use an attendance API to pull daily attendance data into a report.

IoT and Sensor Data:

Internet of Things (IoT) devices like smart watches, water dispensers, or temperature sensors can automatically collect data.

Example:

To study water drinking habits, schools can place smart water meters near fountains.

TIDBIT

IoT devices are useful in Health, safety, and environmental monitoring.

Web Scrapping:

It is a process of programmatically collecting public data from websites. For example, A business student may scrape or extract prices of snacks from online stores to compare them with school canteen prices.

TIDBIT

Web scraping should only be done with teacher for websites that allow it.

Q.6. Explain choosing the right data source and assessing data quality and bias?

Ans: Choosing the Right Data Source;

Not all data is useful. We must choose reliable, accurate, and relevant sources. When selecting a data source, ask a few questions like (i) Is this data from a trusted place? (ii) Is it recent or outdated? and (iii) Does it match the problem I want to solve?

Example: If we want to know favorite snacks of students in 2025, we should avoid using a survey done in 2020.

Assessing Data Quality and Bias:

Data Quality:

Data quality means the data is:

- (i) Complete (no missing values)
- (ii) Correct (no mistakes)
- (iii) Relevant (matches our problem).

Bias

Bias means the data only shows part of the truth and misses the full picture.

Example:

If we only survey boys about their favorite sports, we will miss the opinions of girls. The results will be unfair or biased toward boys.

Measures to avoid Bias:

Remember the following to avoid bias:

- Ask from a diverse group of people.
- Keep questions neutral and not leading.
- Review answers for strange patterns.

TIDBIT

Always collect fresh and problem-related data.

REMEMBER

1. After defining a problem, we need to collect data that can help in solving the problem.
2. There are different methods of data collection such as surveys, APIs, sensors, and web scraping.
3. We should always check for data quality and fairness (bias) before moving to the next step.
4. Data that is incomplete or biased can lead to wrong decisions.

Q.7. Explain Cleaning and Validating Data?

Ans:

After collecting data, the next important step is to clean and validate it. Remember that earlier we collected data using surveys and other methods. Now, we need to check if that data is correct, complete, and ready for use. This step ensures that the data we analyze later is trustworthy.

Raw Data:

Raw data is the original information we collect from people, sensors, websites, or apps. It has not been checked or cleaned yet.

Example: Assume that 20 students filled your survey about favorite sports. Some of the entries are miss or written wrong.

Common Issues in Raw Data

ID.	Error Type	Example
1	Missing data	Blank cells (e.g., no sport name or study time of day)
2	Duplicate entries	Same student entered twice
3	Inconsistent formats	"Cricket", "cricket", "CRICKET"
4	Spelling mistakes	"Football", "Daliy"
5	Outliers (strange values)	Frequency = 1000 times/week

Data Validation Techniques:

Data validation means checking if the data:

- (i) Follows correct rules (e.g., only valid sport names)
- (ii) Has the right type of values (e.g., numbers where numbers are expected)
- (iii) Is not fake or silly

Example Validation Rule:

Only allow one of these values for "Favorite Sport": Football, Cricket, Badminton, Hockey.

Data Cleaning Methods:

Once we find mistakes, we need to fix or remove them. This is called data cleaning.

The following are a few basic steps to clean data:

- Fix spelling errors: "Footbell" → "Football"
- Standardize formats: "CRICKET" → "Cricket"
- Remove duplicates (redundancy): Delete repeated entries for the same student.
- Handle missing values: Ask the person again OR remove the row.
- Fix invalid data: If frequency is written as "ten hundred", fix it or remove it.

Q.8. Explain Analyzing Data? Describe simple tools for analysis?

Ans:

Now that our data is collected, and cleaned properly, we can move on to the analysis steps. This is where we look at the data to find patterns, trends, and answers to the questions. Analyzing data means carefully examining data to (i) Summarize it, (ii) Compare different groups, (iii) Find patterns.

Using Simple Tools for Analysis;

You can start analysing data with simple tools such as Google Sheets, Microsoft Excel or manual writing in a notebook.

Example: You surveyed 20 students. Table 5.4 shows what you may have found.

Favourite Sport	Frequency
Cricket	10
Football	5
Hockey	4
Badminton	1

From this data, we can observe that most students prefer cricket.

REMEMBER

1. Data analysis helps us find patterns and make decisions.
2. We use counts and comparisons to understand data.
3. Even simple tools like tables or spreadsheets can help us analyze data

Q.9. Explain Data Interpreting? Also explain data visualization and communication results?

Ans:

This step is to finding meaning from data (i.e., identifying trends and patterns).

After summarizing the data, we may ask ourselves:

- What do the numbers tell us?
- Why are some options more popular?
- Are there any surprises or patterns?

Example: If we find that most late students live far from the school, we can recommend providing transport support as a solution.

Visualizing and Communicating Results:

People usually understand pictures better than numbers, which is why we use data visualizations.

Remember that in previous steps, we found that cricket was the most popular sport. In this step, we will create a chart that shows this clearly. This makes it easy to present our findings to classmates, teachers, or others.

Importance of Data Visualization:

Data Visualization is important because it

- (i) makes data easier to understand,
- (ii) shows patterns at a glance,
- (iii) helps share ideas clearly in presentations or reports.

Example: Instead of saying "10 students like cricket," we can show it in a bar chart.

Common Types of Charts and Graphs (Bar, Pie, Line, etc.)

Always choose a chart that matches your data. Table 5.6 shows the commonly used types of charts, its use case with examples.

Chart / Bar Type	Used For	Example
Bar Chart	Comparing groups	Number of students who like each sport
Pie Chart	Showing parts of a whole	Percentage of favorite snacks
Line Chart	Showing changes over time	Number of students arriving late each week

TIDBIT

Keep your charts simple, clean, and properly labeled.

Communicating Results:

Once the chart is ready, explain your findings. You can write 2-3 lines:

"Most students (10) like cricket, followed by football (6). Very few students like badminton (1)."

TIDBIT

Use visuals and words together to tell a data story (clear, short, and useful).

REMEMBER

1. Visualizing data means using charts or graphs to show findings.
2. It makes data easier to understand and share.
3. Use the right chart for your data type.
4. Always label and explain your visuals.

Q.10. Describe tools for visualization? Write down the procedure to draw a chart?

Ans:

Once you know which type of chart to use, the next step is to create it using tools. In this step, we create charts make it easier to understand and present data. It is easy to create chart and graphs using specific softwares.

Google Sheets and **Microsoft Excel** are simple yet powerful tools to create charts and graphs from tables.

Steps to Create a Chart in Google Sheets:

- Enter your data into rows and columns.
- Select the data range.
- Click on the "Insert" menu ? choose "Chart."
- Pick the chart type (bar, pie, line).
- Customize the chart with a title, label, and color.

Advantages of Spreadsheet software:

We make use of spreadsheets because they are:

- Easy to use
- No installation needed (for Google Sheets)
- Good for small to medium-sized projects.

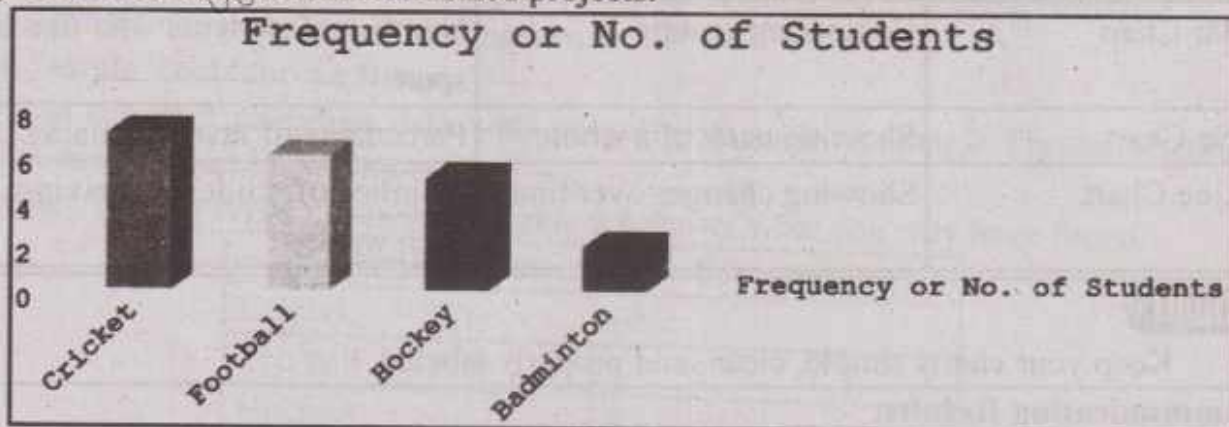


Figure 5.1: Favorite Sport of Students (Based on Data of Table 5.5)

REMEMBER

- Google Sheets and Excel are easy tools for creating charts.
- A dashboard shows multiple visuals in one place to tell a clear data story.

Use of DSLC in real life:

The standard DSLC consists of 6 main steps, however, in some real-world applications/ scenarios, some additional steps may be added. One such step can be Storing and Organizing Data, especially in large projects or team settings where databases and structured storage are important. These are usually managed through systems like DBMS and SQL.

Q.11. Define Database? Why we need structured data? Differentiate between spreadsheet and database?

Ans:

A Database is an organized collection of data. It stores data in the form of tables. Each table is like a mini spreadsheet with rows and columns. Databases help in storing, retrieving, and updating data efficiently.

Purpose of Database:

After cleaning a messy survey table, there is a need to store the cleaned data in a proper format (either in a spreadsheet or a database), so that it is organized, searchable, and ready to analyze. There are two main ways to store structured data i.e.,

- (i) Spreadsheets
- (ii) Databases.

Difference between Database and Spreadsheet:

Method	Description	Example Tool
Spreadsheets	A grid of rows and columns for small datasets	Excel, Google Sheets
Databases	A structured system to store and manage large data in tables	MySQL, SQLite

Database Terms: Meanings and Examples:

Term	Meaning	Example
Table	A group of related data	Sports Survey
Record	A single entry (row) in the table	Ali's full row of data
Field	A specific piece of information	Name, Sport, Frequency

TIDBIT

Spreadsheets are good for small projects. Databases are better for big or multi-user data systems.

Q.12. Define DBMS? Also explain some queries written in SQL?

Ans:

A Database Management System (DBMS) is software used to create, manage, and interact with databases. Examples include MySQL, SQLite, and Microsoft Access.

To work with data in a database, we use a language called Structured Query Language (SQL). SQL helps us to add, update, delete, or retrieve data using simple commands. Some basic SQL queries are as follows:

- **SELECT * FROM Sports:** This simple query shows all the data from the table named Sports.
- **SELECT Name FROM Sports WHERE Favorite_Sport = 'Cricket':** This query shows names of students present in Sports table who like cricket.

- **INSERT INTO Sports (Name, Favorite_Sport) VALUES ('Ali', 'Football');**
This query adds a new entry.

These basic SQL statements help in managing and retrieving specific data from large databases efficiently.

REMEMBER

1. Data should be stored in a structured way for easy access and analysis.
2. **Spreadsheets** are useful for small datasets.
3. **Databases** are used for bigger and more complex data.
4. A **table** stores data using **records (rows)** and **fields (columns)**.
5. Organized data reduces duplication errors and makes analysis easier.

Objective Type Questions

Additional MCQ's

Multiple Choice Questions prepared in the light of new Examination Techniques
(Knowledge, Understanding, Application, Analytical & Conceptual)

Introduction

1. **Purpose of data Science:**
(A) Represent data in Formal way (B) Convert data into information
(C) Both a, b (D) None of them
2. **Which of the following is a method of data collection?**
(A) Surveys (B) Social Media (C) Mobile Apps (D) All above
3. **Which of the following is not Advantage of data Science?**
(A) Predicting weather (B) Calculate cash
(C) Health services (D) Customer care
4. **Why organization use data Science?**
(A) Decision Making (B) Product Development
(C) Control Machines (D) None of them

5.1

Introduction to Data Science

5. **Which of the following data is not useful for data Science?**
(A) Fitness tracking Data (B) Google Search
(C) Shopping online (D) None of them
6. **Which of the following is a data science technique?**
(A) organizing (B) Analyzing (C) Visualizing (D) All above
7. **Which can lead to mistakes?**
(A) Analyzing data (B) Make a guess
(C) Visualizing (D) Collection through survey

8. Suggestion of products is done by _____ platform:
 (A) Social Media (B) Health Care
 (C) Transport (D) Online Shopping
9. Which of the following is not an application of data Science?
 (A) Education (B) Transport
 (C) collection of currency notes (D) Health Care
10. Which Data Science application is used to track progress of learning improvement?
 (A) Education (B) Transport (C) Health Care (D) online shopping

5.2

Overview of Data Science Life Cycle

11. DSLC stands for:
 (A) Data Science long cycle (B) Data Science large cycle
 (C) Data Science life cycle (D) Data Science life circle
12. Total number of steps in DSLC:
 (A) 2 (B) 4 (C) 6 (D) 8
13. Data Science life cycle is helpful in:
 (A) Data collection (B) Problem solving (C) Surveys (D) Analyzing
14. Problem understanding means:
 (A) collecting Data (B) Data Analyzing
 (C) Importance of Problem (D) None of them
15. Making a clear target to solve a problem is referred to:
 (A) Understanding Data (B) collecting Data
 (C) Analyzing Data (D) Interpreting Data
16. When we refine a problem, we _____:
 (A) use clear words (B) Focus on target (C) Avoid unclear (D) All above
17. Importance to setting clear objective:
 (A) calculate cost (B) Guiding data Science process
 (C) show schedule (D) Defining Targets
18. Which of the following allow user to focus on Project?
 (A) Identity Problem (B) Setting clear object
 (C) Both a, b (D) None of them
19. Next step after Problem identification:
 (A) Data collection (B) Data cleaning
 (C) Data Analyzing (D) Data visualizing
20. Method in which we asked question and record answers:
 (A) IoT and sensor data (B) APIs source
 (C) Surveys (D) Web scraping
21. Which of the following is the best tool for conduct online survey?
 (A) Social Media (B) Google Form (C) MS Word (D) Google colab

22. Which allow software to collect data from websites or applications?
(A) Surveys (B) IoT & Sensor Data
(C) APIs (D) Web scraping
23. IoT is useful in:
(A) Transportation (B) Health and safety
(C) Education (D) online shopping
24. Public data collection method is:
(A) Surveys (B) IoT & Sensors (C) Web scraping (D) APIs
25. Data only shows part of truth and miss the full picture:
(A) Incomplete (B) Bias (C) Reliable (D) relevant
26. Raw data is often:
(A) Messy (B) Incomplete (C) Irrelevant (D) All above
27. Data Validation is important for:
(A) Cleaning (B) Arrangement (C) Analysis (D) Relevant
28. Data which has not been checked or cleaned:
(A) Raw Data (B) Reliable Data (C) Relevant Data (D) Complete Data
29. How many types of common Data Errors?
(A) 1 (B) 2 (C) 3 (D) 4
30. How many type of Data validation techniques are there?
(A) 1 (B) 2 (C) 3 (D) 4
31. _____ tool help to validate data:
(A) Ms word (B) Ms-Excel (C) Google sheet (D) Both b, c
32. Redundancy handling is done in:
(A) Data Collection (B) Data Analyzing
(C) Data cleaning (D) Data visualizing
33. After data cleaning next step is:
(A) Data Analysis (B) Data interpretation
(C) Data visualization (D) None of them
34. Find trends & patterns involved in:
(A) Data Analysis (B) Data interpretation
(C) Data visualization (D) Data cleaning
35. Which of the following is used as a tool for Analysis?
(A) MS Word (B) CorelDraw (C) Google sheet (D) Google Form
36. Which method is used to find meaning from data?
(A) Data Analysis (B) Data Interpreting
(C) Data visualization (D) None of them
37. Final step of DSLC method:
(A) Data visualization (B) Data Interpretation
(C) Data Analysis (D) None of them
38. Share the results is called:
(A) Interpreting (B) Communicating (C) Presenting (D) Explanation

39. Data visualization tool is:
 (A) CorelDraw (B) Photoshop (C) MS Excel (D) MS Word
40. data visualization tools helps in:
 (A) DATA Analysis (B) Data Representation
 (C) Decision making (D) All above
41. Data in charts and graphs are inserted through:
 (A) Tables (B) images (C) Paragraphs (D) Reports
42. Histogram is a:
 (A) Graph (B) Chart (C) Both a, b (D) None of them

43. Storing and organizing Data is called:
 (A) Data Gathering (B) Database (C) Data collection (D) storage
44. SQL is a:
 (A) Survey software (B) Data Collection software
 (C) Database Language (D) None of them
45. Which of the following is a best tool to store data?
 (A) MS Excel (B) Google sheet (C) Database (D) Google Form
46. Which Structure system store large Data?
 (A) Spreadsheet (B) Google sheet (C) MySQL (D) All above
47. Database store Data in the form of:
 (A) Reports (B) Tables (C) images (D) Paragraphs
48. MS Access is a:
 (A) Database (B) Data Structure (C) DBMS (D) SQL
49. Which of the following is a DBMS?
 (A) MS Excel (B) MS-Word (C) MS-Powerpoint (D) MySQL
50. In DBMS, Rows are also called:
 (A) Fields (B) columns (C) cells (D) Records

Answers

- | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|
| 1. C | 2. D | 3. B | 4. A | 5. D | 6. D | 7. B |
| 8. D | 9. C | 10. A | 11. C | 12. C | 13. B | 14. C |
| 15. A | 16. D | 17. B | 18. B | 19. A | 20. C | 21. B |
| 22. C | 23. B | 24. C | 25. B | 26. D | 27. C | 28. A |
| 29. C | 30. C | 31. D | 32. C | 33. A | 34. A | 35. C |
| 36. B | 37. A | 38. B | 39. C | 40. D | 41. A | 42. A |
| 43. B | 44. C | 45. C | 46. C | 47. B | 48. C | 49. D |
| 50. D | | | | | | |

Additional Short Question's

Short Answer Questions prepared in the light of new Examination Techniques
(Knowledge, Understanding, Application, Analytical & Conceptual)

Introduction

1. Define Data Science.

Ans: Data science is the process of using data to understand problems, discover patterns, and make smart decisions.

Example: Daraz online shopping app use data science to increase their sales.

2. Why data science is important?

Ans: Data Science helps us turn this raw data into useful information. Whether it is predicting weather, understanding interests of students, or improving healthcare. Data Science is transforming the way we live, learn, and work.

3. How data science techniques used by business organizations?

Ans: Organizations across the whole world use data to improve:

- Decision-making
- Save costs
- Create better services.

4. Write down the names of steps of DSLC.

Ans: Data Science life Cycle include:

- Understanding the problem
- Collecting and cleaning data
- Analysing Data
- Interpreting data
- Visualizing and communicating the results

5.1

Introduction to Data Science

5. What is the purpose of data science?

Ans: Data Science is the field where we use data to find patterns, answer questions, and help people or businesses make better decisions. It involves collecting, organizing, analyzing, and visualizing data.

6. Why data is important in decision making?

Ans: Data plays a big role in helping people and organizations make better decisions. Without data, decisions may be based on guesses or personal opinions which can lead to mistakes.

Example:

if the school administration desires to start a new sports club, they can first collect data on which sports students enjoy most. This helps them make a smart and fair choice.

7. **Explain real life applications of data science in social media and online shopping.**

Ans: Social media:

Applications like Facebook, YouTube and Instagram use data to show us posts and videos we might like

Online shopping:

E-commerce websites suggest products using data of our previous purchases.

8. **Explain real life applications of data science in health science and education.**

Ans: Healthcare:

Hospitals use data science to predict diseases and manage patient care more effectively.

Education:

In schools, data from student attendance and test results is used to improve learning strategies and track progress.

5.2

Overview of Data Science Life Cycle

9. **Define DSLC and write the name of its steps.**

Ans: Data Science is not covered in a single step. It follows a series of steps that are helpful in solving a problem. This is what we call Data Science Life Cycle (DSLC). DSLC usually comprises of the following steps:

- Understanding the problem
- Collecting Data
- Cleaning and Validating data
- Analyzing Data
- Interpreting Data
- Visualizing and Communication Results

10. **Why it is important to understand the problem?**

Ans: It is necessary to clearly understand the problem under consideration. In case, we do not understand the problem properly, we may collect the wrong data resulting in useless or wrong results.

11. **How many steps are considered to identify the problem?**

Ans: When defining a problem, we need to consider the following:

- (i) Use clear and simple words
- (ii) Focus on one specific question
- (iii) Avoid unclear or general terms.

12. **What is meant by setting the clear goal?**

Ans: Once we have identified and defined the problem, we must decide what we want to achieve. These are called our objectives which help in guiding the rest of the data science process. They state:

- (i) What kind of data we need
- (ii) What questions we will ask
- (iii) What results we are looking for.

13. What are the advantages of clear understand a problem?

Ans:

- Good defining problems are solvable using data.
- Clear objectives help in collecting the right data and reaching useful results.

14. Define data collection and name the steps of data collection methods.

Ans: Data is like raw material, and we must gather the right kind of data before we can clean, store, or analyze it.

Data Collection Methods:

- Survey
- APIs and online source
- IoT and Sensors
- Web scraping

15. Define surveys with example.

Ans: Surveys are the easiest and most common method. We ask people questions and record their answers.

Example:

To find out which sports students enjoy, we can create a survey with questions using Google form and share its link with students for response.

16. What is the purpose of APIs and online sources?

Ans: Application Programmer Interface (APIs) allow software collect data from websites or applications (apps in short) automatically.

Example:

A school may use an attendance API to pull daily attendance data into a report.

17. Define IoT and Sensor data.

Ans: Internet of Things (IoT) devices like smart watches, water dispensers, or temperature sensors can automatically collect data.

Example:

To study water drinking habits, schools can place smart water meters near fountains

18. What is web scraping?

Ans: It is a process of programmatically collecting public data from websites. For example, A business student may scrape or extract prices of snacks from online stores to compare them with school canteen prices.

19. What is data quality?

Ans: Data quality means the data is:

- (i) Complete (no missing values)
- (ii) Correct (no mistakes)
- (iii) Relevant (matches our problem).

20. What is bias in data collection?

Ans: Bias means the data only shows part of the truth and misses the full picture.

Example:

If we only survey boys about their favorite sports, we will miss the opinions of girls. The results will be unfair or biased toward boys.

21. What measures are taken to avoid bias?

Ans: Remember the following to avoid bias:

- Ask from a diverse group of people.
- Keep questions neutral and not leading.
- Review answers for strange patterns.

22. What is raw data? Give an example.

Ans: Raw data is the original information we collect from people, sensors, websites, or apps. It has not been checked or cleaned yet.

Example:

Assume that 20 students filled your survey about favorite sports. Some of the entries are miss or written wrong.

23. Describe data validation techniques.

Ans: Data validation means checking if the data:

- (i) Follows correct rules (e.g., only valid sport names)
- (ii) Has the right type of values (e.g., numbers where numbers are expected)
- (iii) Is not fake or silly

24. Define data cleaning methods.

Ans: Once we find mistakes, we need to fix or remove them. This is called data cleaning.

The following are a few basic steps to clean data:

- Fix spelling errors: "Footbell" → "Football"
- Standardize formats: "CRICKET" → "Cricket"
- Remove duplicates (redundancy): Delete repeated entries for the same student.
- Handle missing values: Ask the person again OR remove the row.
- Fix invalid data: If frequency is written as "ten hundred", fix it or remove it.

25. Name the simple tools for analyzing data.

Ans: You can start analysing data with simple tools such as Google Sheets, Microsoft Excel or manual writing in a notebook.

26. Explain communication the results.

Ans: People usually understand pictures better than numbers, which is why we use data visualizations.

Remember that in previous steps, we found that cricket was the most popular sport. In this step, we will create a chart that shows this clearly. This makes it easy to present our findings to classmates, teachers, or others.

27. What is the importance of data visualization?

Ans: Data Visualization is important because it

- (i) makes data easier to understand,
- (ii) shows patterns at a glance,
- (iii) helps share ideas clearly in presentations or reports.

28. Write down the steps to create a chart using google sheet.

Ans: Steps to Create a Chart in Google Sheets:

- i. Enter your data into rows and columns.
- ii. Select the data range.
- iii. Click on the "Insert" menu → choose "Chart."
- iv. Pick the chart type (bar, pie, line).
- v. Customize the chart with a title, label, and color.

29. What are advantages of spreadsheet software?

Ans: We make use of spreadsheets because they are:

- (i) Easy to use
- (ii) No installation needed (for Google Sheets)
- (iii) Good for small to medium-sized projects.

30. Describe the uses of DSLC in real life.

Ans: The standard DSLC consists of 6 main steps, however, in some real-world applications/ scenarios, some additional steps may be added. One such step can be Storing and Organizing Data, especially in large projects or team settings where databases and structured storage are important. These are usually managed through systems like DBMS and SQL.

31. Define Database.

Ans: A Database is an organized collection of data. It stores data in the form of tables. Each table is like a mini spreadsheet with rows and columns. Databases help in storing, retrieving, and updating data efficiently.

32. What is the purpose of Database?

Ans: Purpose of Database:

After cleaning a messy survey table, there is a need to store the cleaned data in a proper format (either in a spreadsheet or a database), so that it is organized, searchable, and ready to analyze. There are two main ways to store structured data i.e.,

- (i) Spreadsheets
- (ii) Databases.

33. Define DBMS. Give example.

Ans: A Database Management System (DBMS) is software used to create, manage, and interact with databases.

Examples:

MySQL, SQLite, and Microsoft Access.

34. Explain some queries used in SQL.

Ans: Some basic SQL queries are as follows:

- **SELECT * FROM Sports:** This simple query shows all the data from the table named Sports.
- **SELECT Name FROM Sports WHERE Favorite_Sport = 'Cricket':** This query shows names of students present in Sports table who like cricket.
- **INSERT INTO Sports (Name, Favorite_Sport) VALUES ('Ali', 'Football');** This query adds a new entry. These basic SQL statements help in managing and retrieving specific data from large databases efficiently.

Exercise Questions (Solved)

Multiple Choice Questions

- What is the first step in the Data Science Life Cycle (DSLCL)?
 - Collecting Data
 - Visualizing Data
 - Understanding the Problem
 - Analyzing the Data
- Which method is best for collecting opinions from many students?
 - IoT Devices
 - Survey
 - Web Scraping
 - Database
- What is "bias" in data?
 - Very large data size
 - Unfair or one-sided data
 - Well-cleaned data
 - High-speed data
- Which tool is ideal for beginners to create charts?
 - Tableau
 - Python
 - Google Sheets
 - Hadoop
- A relational store information in the form of:
 - Slides
 - Tables
 - Blocks
 - Layers
- Which of the following charts is best for comparing groups?
 - Pie Chart
 - Bar Chart
 - Line Chart
 - Area Chart
- What is the main goal of analyzing data?
 - Delete old data
 - Decorate data
 - Find patterns and make decisions
 - Upload to the internet
- Which of the following tools is an advanced data visualization platform?
 - MS Paint
 - Google Data Studio
 - Notepad
 - WhatsApp
- What should be done before storing data?
 - Add animations
 - Clean and validate it
 - Encrypt it
 - Translate it
- Which one is an example of a data source?
 - YouTube
 - Survey form
 - Whiteboard
 - Slide show

Answers

- | | | | | |
|--------|--------|--------|--------|---------|
| 1. (c) | 2. (b) | 3. (b) | 4. (c) | 5. (b) |
| 6. (b) | 7. (c) | 8. (b) | 9. (b) | 10. (b) |

Short Questions

1. What is the importance of understanding a problem before collecting data?

Ans: Before we start talking about collecting data, it is necessary to clearly understand the problem under consideration. In case, we do not understand the problem properly, we may collect the wrong data resulting in useless or wrong results.

2. Name any two methods of collecting data and give examples of each.

Ans: Survey:

we can create a survey with questions using Google form and share its link with students for response.

APIs: A school may use an attendance API to pull daily attendance data into a report.

3. What do we mean by "cleaning" data?

Ans: Remember that earlier we collected data using surveys and other methods. Now, we need to check if that data is correct, complete, and ready for use. This process is called Data Cleaning. This step ensures that the data we analyze later is trustworthy.

4. Why it is important to validate data before using it?

Ans: Data validation important because it checking if the data:

- (i) Follows correct rules (e.g., only valid sport names)
- (ii) Has the right type of values (e.g., numbers where numbers are expected)
- (iii) Is not fake or silly.

5. What is the purpose to visualizing data using charts?

Ans: People usually understand pictures better than numbers, which is why we use data visualizations. Data Visualization is important because it

- (i) makes data easier to understand,
- (ii) shows patterns at a glance,
- (iii) helps share ideas clearly in presentations or reports.

6. How does bias in data affect results?

Ans: Bias means the data only shows part of the truth and misses the full picture.

Example:

If we only survey boys about their favorite sports, we will miss the opinions of girls. The results will be unfair or biased toward boys.

7. Write two steps you will follow to clean a messy survey.

Ans: The following are few steps to clean data:

- Standardize formats: "CRICKET" → "Cricket"
- Remove duplicates (redundancy): Delete repeated entries for the same student.

8. Name any two tools used for creating charts.

Ans: Google Sheets and Microsoft Excel are simple yet powerful tools to create charts and graphs from tables.

Long Questions

1. Explain the six steps of the Data Science Life Cycle (DSLCL) with one line each.

Ans: Data Science is not covered in a single step. It follows a series of steps that are helpful in solving a problem. This is what we call Data Science Life Cycle (DSLCL). DSLCL usually comprises of the following steps:

Understanding the problem:

It is necessary to clearly understand the problem under consideration. In case, we do not understand the problem properly, we may collect the wrong data

Collecting Data:

Data is like raw material, and we must gather the right kind of data before we can clean, store, or analyze it. There are many ways to collect data, depending on the problem and available tools.

Cleaning and Validating data:

we need to check if that data is correct, complete, and ready for use. This step ensures that the data we analyze later is trustworthy.

Analyzing Data:

Analyzing data means carefully examining data to (i) Summarize it, (ii) Compare different groups, (iii) Find patterns.

Interpreting Data:

This step is to finding meaning from data (i.e., identifying trends and patterns)

Visualizing and Communication Results:

In this step, we will create a chart that shows this clearly. This makes it easy to present our findings to classmates, teachers, or others.

2. Give a detailed comparison between Bar Chart, Pie Chart, and Line Chart with examples.

Ans:

Feature/Aspect	Bar Chart	Pie Chart	Line Chart
Primary Purpose	To compare different categories or groups.	To show the proportion or percentage of a whole (parts of a whole).	To show trends, patterns, or changes over a continuous period (usually time).
Data Type	Categorical or discrete data.	Categorical data, where categories sum up to a whole.	Continuous data, especially time-series data.
Axes	One axis shows categories, the other shows values.	No axes; it is a circle divided into slices.	Both axes show continuous variables (e.g., time on X-axis, value on Y-axis).
Visual Representation	Rectangular bars with lengths proportional to the values they represent.	A circle divided into slices, where the angle or area of each slice is proportional to its quantity.	Data points connected by straight lines.
Best Use Case	Comparing sales figures of different products, population of different cities, etc.	Showing market share of companies, budget allocation, survey percentages, etc.	Showing stock prices over a year, website traffic over a week, temperature changes, etc.

Strengths	Easy to read and compare individual categories. Good for showing exact differences.	Visually simple and effective for showing simple proportions at a glance.	Excellent for displaying trends, patterns, and rates of change over time.
Weaknesses	Can become cluttered if there are too many categories.	Difficult to compare individual slices if there are many small slices. Cannot show trends over time.	Can be misleading if the time intervals are not consistent. Not good for comparing discrete categories.
Example	Comparing the population of five different countries: USA, India, China, Brazil, and Nigeria.	Showing the market share of smartphone operating systems: Android (70%), iOS (28%), Others (2%).	Showing the monthly sales of a company from January to December.

3. Differentiate between DBMS and SQL.

Ans:

Feature/Aspect	DBMS (Database Management System)	SQL (Structured Query Language)
Basic Definition	It is a software system used to create, manage, and organize databases.	It is a programming language used to communicate with and manipulate databases.
Nature	It is a software application or a system (e.g., MySQL, Oracle).	It is a query language or a syntax used to perform tasks.
Function	It acts as an interface between the database and the users or applications.	It acts as a tool to execute specific commands like retrieving, inserting, or updating data.
Components	Includes utilities, administrators, storage managers, and a query processor to run SQL commands.	Consists of commands, clauses, and operators (e.g., SELECT, INSERT, WHERE).
Purpose	To store, retrieve, and manage data efficiently while ensuring security and integrity.	To perform specific operations on the data stored in the DBMS.

4. Write a Query in SQL to get all data of a student who's roll_number is 15 from student table.

Ans:

SELECT

FROM

STUDENT

WHERE

roll_number = 15;