UNIT 5

Data Science

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 (DSLC).
- 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.

Introduction

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. 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 (DSLC), which includes understanding the problem, collecting and cleaning data, analyzing 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, analyze, 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.

5.1 Introduction to Data Science

We are living in a world full of data. Whether we are using a mobile phone, searching on Google, shopping online, or even walking with a fitness tracker, data is being collected. The question arises how is this data useful?

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. For 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.

Real-Life Applications of Data Science

Data Science is used in many aspects of our daily life often without our notice.

- **Social media:** Applications like Facebook, YouTube and Instagram use data to show us posts or video 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.

Data Science supports in making our lives smarter, easier, and more personalized from watching YouTube videos to receiving health tips or choosing the best route.

5.2 Overview of the Data Science Life Cycle

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:

Step 1: Understanding the Problem

Step 2: Collecting Data
Step 3: Cleaning and Validating Data
Step 4: Analyzing Data
Step 5: Interpreting Data
Step 6: Visualizing and Communicating Results

Understanding the Problem

Before we start talking about collecting data, it is necessary to clearly understand the problem under consideration. Understanding the problem is the most important step in DSLC. 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 (what, why, who) such as:

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

The following sub-sections cover step 1 of DSLC in more detail.

Identifying and Defining a Problem

First, we need to **identify a problem** that we want to solve using data. The problem can be related to our school, home, city, or any area of life where decisions are made based on information. Once a problem is identified, then we must **define it clearly and simply**.

Example: Instead of saying, ***"students are not happy,"*** we can define the problem in a better way as ***"Which school activities do students enjoy the most?"***

When defining a problem, we need to consider the following: (i) Use clear and simple words, (ii) focus on one specific question, and (iii) avoid unclear or general terms.

TIDBIT

A good problem is one that can be answered by collecting and analyzing data.

Setting Clear Objectives

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, and (iii) what results we are looking for.

Example: For the problem stated in above "Which school activities do students enjoy the most?", we can formulate our objectives as follows:

- Find out how many students participate in each activity.
- Find out which activity is rated as most enjoyable by the students.
- Recommend which activities should be improved or replaced.

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.
		Recommend steps to promote healthy habits.
Business	*A school canteen wants to know which snacks students like the most.*	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. Clear problems are solvable using data.
4. Clear objectives help in collecting the right data and reaching useful results.

Collecting Data

Once we understand the problem and set clear goals, the next step in DSLC is to collect the required data. Data is like raw material, and we must gather the right kind of data before we can clean, store, or analyze it.

In **Step 1**, we asked questions like ***"Which school activities do students enjoy the most?"*** or ***"Why are students arriving late to school?"***

Step 2 is about collecting information that can help us answer these questions.

We will explore:

(i) data collection methods

(ii) where to collect data from (sources)

(iii) How to make sure it is useful and fair (data quality)

(i) Methods of Data Collection: There are many ways to collect data, depending on the problem and available tools. Let us discuss the most common ones used in schools, homes, and businesses.

- **Surveys (Online/paper based):** Surveys are the easiest and most common method. We ask people questions and record their answers. **For example:** To find out which sports students enjoy, we can create a survey with questions such as (i) Which sport do you like the most? (ii) How often do you play it? (iii) What time of day do you prefer to play?

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. **For example**, A school may use an attendance API to pull daily attendance data into a report (use more by developers).
- **IoT and Sensor Data:** Internet of Things (IoT) devices like smart watches,

water dispensers, or temperature sensors can automatically collect data. **For 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 Scraping:** It is a process of programmatically collecting public data from websites. **For example**, A business student may scrape 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.

(ii) 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.

TIDBIT

Always collect fresh and problem-related data.

(iii) Assessing Data Quality and Bias

Data quality means the data is (i) Complete (no missing values), (ii) Correct (no mistakes), and (iii) Relevant (matches our problem).

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

Remember the following to avoid bias:

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

ACTIVITY

Designing a Simple Survey

Objective: Learn how to collect and store data by creating a short survey about free time habits of your classmates.

Instructions:

1. **Design Your Survey:** Think of **3 to 5 questions** to ask your classmates.
2. **Prepare to Collect Responses:** Choose a method (e.g., paper based or Google online form)
3. **Collect Data:** Ask at least 5 classmates to answer your survey and write down their responses clearly.
4. **Store the Data:** (i) If you are using paper, keep it neat and organized in columns. (ii) If using digital tools, save the file with a clear name like ***FreeTime_YourName.xlsx***
5. **Optionally**, show your teacher or classmates how you collected and stored the 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.

Cleaning and Validating Data

After collecting data, the next important step is to clean and validate it. Raw data is often messy. It may contain errors, missing values, or even wrong entries. If we use this data without checking it, our analysis will give wrong results. 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. For example, assume that 20 students filled your survey about favorite sports. Some entries may look like as:

Table 5.2: Raw Data about Favorite Sport of Students

ID.	Name	Favorite Sport	Frequency	Study Time of Day
1	Ahmad	cricket	Daily	Evening
2	Ali		Weekly	Afternoon
3	Sara	Footbal	Daily	
4	Sana	Football	Daliy	Evening

Table 5.2 contains:
(i) spelling errors i.e., Footbal (ID 3), Daliy (ID 4)
(ii) missing entries (ID 2 & 3)
(iii) Mixed Capitalizations i.e., cricket (ID 1), Football (ID 4). Such kind of data needs cleaning.

Common Data Errors (Missing, Duplicate, Outliers)

Table 5.3 shows some common issues found in raw data.

Table 5.3: 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	"Footbal", "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.

Validation in Surveys: For this purpose, consider the following:

- Use multiple choice instead of blank text boxes.
- Make important questions as required, so no one skips them.

- Use dropdowns or checkboxes to reduce spelling mistakes.

TIDBIT

Tools like **Google Sheets**, **Excel**, provides rules that can help validate data.

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: "Footbal" → "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.

Why Cleaning Data Matters

Suppose that we skip this step and we analyze data that includes:

- Misspelled sports ("Cricket")
- Missing Times of the Day / Study time.
- Duplicates

Our result might say: "Football: 3 votes, cricket: 1 vote, Cricket: 2 votes" → This gives the wrong picture, because "cricket" and "Cricket" are the same sport!

In summary, we can say that clean data leads to smart decisions. Dirty data leads to incorrect result.

Analyzing Data

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 we asked in Step 1.

Suppose that you have a clean and organized table of data, for example, student responses about their favorite sport or study times. Now you may ask question like "what does this data tell us?" This step helps us make decisions based on

what we discover in the data.

What Does it mean to Analyze Data

Analyzing data means carefully examining data to (i) Summarize it (e.g., how many students chose each option), (ii) Compare different groups (e.g., boys vs girls, morning vs evening), (iii) Find patterns (e.g., most students prefer football and play in the evening).

You may use measures like counts (how many?), averages, frequency (how often?) etc.

Using Simple Tools for Analysis

You can start analyzing 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.

Table 5.4: Data Derived from Student Survey

Favorite Sport	Frequency
Cricket	10
Football	5
Hockey	4
Badminton	1

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

ACTIVITY Data Cleaning

Objective: Learn how to identify and correct common data entry mistakes (such as spelling errors, missing values, and inconsistent formatting).

Instructions:

1. Review the Messy Table 5.5

Table 5.5: Raw Data about Favorite Sport of Students (with Messy Data)

ID	Name	Favorite Sport	Frequency	Time of Day
1	Sadiq	cricket	Daliy	Evening
2	Noor	Cricket	Weekly	
3	Maryam	Footbal		morning
4	Zara	Football	Daliy	Afternoon

2. Work in pairs: Sit with a partner and study the table together.

3. Spot and List Errors:

Find **at least 5 errors**. These may include:

- Spelling mistakes (e.g., *Cricket, Daily*)
- Missing data (e.g., blank cells)
- Inconsistent formatting (e.g., *Football* vs *football*, *morning* not capitalized)

4. Correct the Errors:
(i) Discuss how to fix each error logically and fairly.
(ii) Rewrite the **cleaned table** neatly in your notebook.
5. Class Sharing (optional):
(i) Compare your cleaned table with another group.
(ii) See if they found the same errors or different ones.

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.

ACTIVITY

Data Analysis

Objective: Learn how to analyze data by identifying trends and drawing basic conclusions from a frequency table.
Instructions:

- Use your own survey data (if available) or the sample data given in Table 5.6 below:

Table 5.6: Sample Derived Data from Survey Data

Sport	Frequency or No. of Students
Cricket	7
Football	6
Hockey	5
Badminton	2

- Carefully read the table showing how many students like each sport.
- Identify: (i) The **most popular** sport, (ii) The **least popular** sport
- Write **one reason** why each sport might be most or least popular in your class.
- Present your findings in your notebook using simple sentences.

Interpreting Data

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

After analyzing the data, the final step is to share the results in a way that others can understand easily. People usually understand pictures better than numbers, which is why we use data visualizations.

Remember that in previous step, 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 5.7 shows the commonly used types of charts, its use case with examples.

Table 5.7: Common Types of Charts

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.

5.3 Tools for Visualization

Once you know which type of chart to use, the next step is to create it using tools. In Step 5, we learned how charts make it easier to understand and present data. In this section, you will learn **how to actually make those charts** using software. Let us explore the most useful and beginner-friendly tools.

Creating Charts in Spreadsheets (Excel/Google Sheets)

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:

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.

Example: Assume that you have data (refer to Table 5.6) in Google Sheets. You are encouraged to use these steps to turn Table 5.6 data into a bar chart.

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 Figure 5.1 shows a simple bar graph corresponding to our Table 5.6.

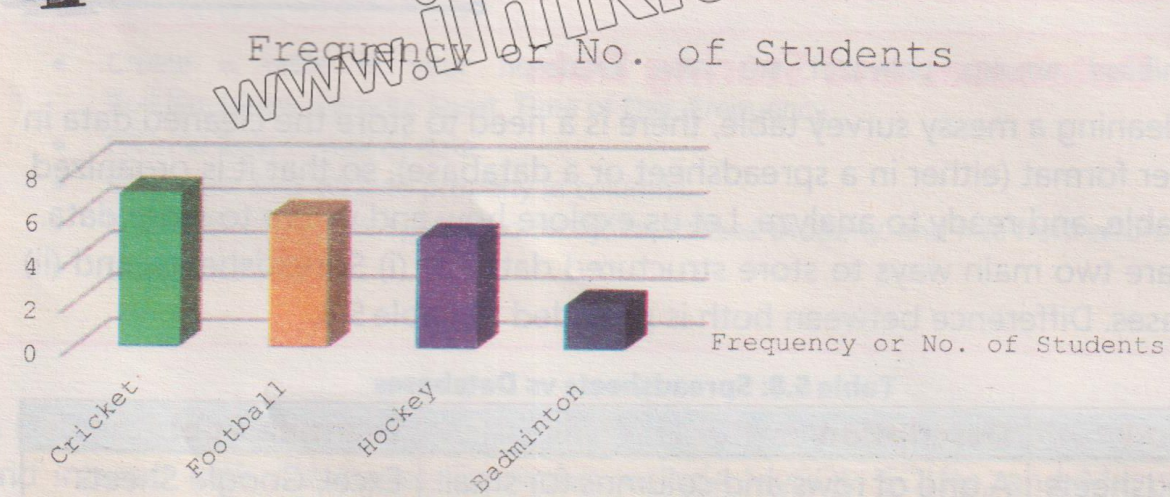

Figure 5.1: Favorite Sport of Students (Based on Data of Table 5.6)

ACTIVITY

Drawing a Bar Cart, Pie chart and Line Chart from Survey Data given in table 5.6

Objective: Learn how to convert a frequency table into a bar chart to visually represent survey data.

Instructions:

- Use the sample data given in **Table 5.6**.
- Draw a **bar chart, pie chart and line chart** in your notebook.
- Present your findings in your notebook using simple sentences.

REMEMBER

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

Note: 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.

5.4. Databases and Storing Data

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. Let us explore how and where to store data. There are two main ways to store structured data i.e., (i) Spreadsheets, and (ii) Databases. Difference between both is provided in Table 5.8.

Table 5.8: Spreadsheets vs Databases

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

TIDBIT

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

What is Database?

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. Table 5.9 breaks down database related terms.

Table 5.9 : 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

ACTIVITY

Data Organization

Objective: Learn how to organize cleaned data into a well-structured table using consistent formatting.

Instructions:

- Use the cleaned survey data you collected earlier.

- Create a table in your notebook with the following column headings: Roll No, Name, Favorite Sport, Time of Day, Frequency
- Fill in at least 5 rows of data.
- Keep your data aligned properly in columns.
- Ensure all entries are spelled correctly, capitalized properly, and free from errors.
- DBMS and SQL.

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 errors and makes analysis easier.

Summary

- **Data Science** is the process of using data to understand problems, identify patterns, and make informed smart decisions.
- **The Life Cycle (DSLC):** It follows a 6-step cycle: Problem Understanding, Data Collection, Cleaning/Validation, Analysis, Interpretation, and Visualization.
- **Data Collection:** Data is gathered from various sources like surveys (Google Forms), APIs, web scraping, and IoT sensors.
- **Data Quality:** Reliable data must be accurate, recent, and unbiased; diverse groups must be surveyed to avoid misleading results.
- **Cleaning & Validation:** "Messy" raw data must be cleaned by removing duplicates and fixing errors to ensure the results are trustworthy.
- **Data Visualization:** Charts turn complex numbers into easy-to-read visuals; **Bar Charts** compare groups, **Pie Charts** show parts of a whole, and **Line Charts** show trends over time.
- **SQL Language:** Structured Query Language (SQL) is used to communicate with databases using commands like **SELECT** (to get data) and **INSERT** (to add data).

EXERCISE

Multiple Choice Questions:

1. **What is the first step in the Data Science Life Cycle (DSLC)?**

   (a) Collecting Data (b) Visualizing Data
   (c) Understanding the Problem (d) Analyzing the Data

2. **Which method is best for collecting opinions from many students?**

   (a) IoT Devices (b) Survey
   (c) Web Scraping (d) Database

3. **What is "bias" in data?**

   (a) Very large data size (b) Unfair or one-sided data
   (c) Well-cleaned data (d) High-speed data

4. **Which tool is ideal for beginners to create charts?**

(a) Tableau (b) Python
(c) Google Sheets (d) Hadoop

5. **A relational database store information in the form of:**

(a) Slides (b) Tables
(c) Blocks (d) Layers

6. **Which of the following charts is best for comparing groups?**

(a) Pie Chart (b) Bar Chart
(c) Line Chart (d) Area Chart

7. **What is the main goal of analyzing data?**

(a) Delete old data (b) Decorate data
(c) Find patterns and make decisions (d) Upload to the internet

8. **Which of the following tools is an advanced data visualization platform?**

(a) MS Paint (b) Google Data Studio
(c) Notepad (d) WhatsApp

9. **What should be done before storing data?**

(a) Add animations (b) Clean and validate it
(c) Encrypt it (d) Translate it

10. **Which one is an example of a data source?**

(a) YouTube (b) Survey form
(c) Whiteboard (d) Slide show

Short Questions

1. What is the importance of understanding a problem before collecting data?
2. Name any two methods of collecting data and give one example of each.
3. What do we mean by "cleaning" data?
4. Why is it important to validate data before using it?
5. What is the purpose of visualizing data using charts?
6. How does bias in data affect results?

7. Write two steps you will follow to clean a messy survey.
8. Name any two tools used for creating charts.

Long Questions:

1. Explain the six steps of the Data Science Life Cycle (DSLC) with one line each.
2. Give a detailed comparison between Bar Chart, Pie Chart, and Line Chart with examples.
3. Differentiate between DBMS and SQL
4. Write a Query in SQL to get all data of a student who's roll_number is 15 from student table.

Answer Key for Multiple Choice Questions

1. **c)** - Understanding the Problem
2. **b)** - Survey
3. **b)** - Unfair or one-sided data
4. **c)** - Google Sheets
5. **b)** - Tables
6. **b)** - Bar Chart
7. **c)** - Find patterns and make decisions
8. **b)** - Google Data Studio
9. **b)** - Clean and validate it
10. **b)** - Survey form