CLASS XI • COMPUTER SCIENCE • STUDY BOOKLET

# Unit 5: Data Analytics

*Complete Question Bank — MCQs, Short Questions, Long Questions & Solved Textbook Exercise*

50 MCQs | 40 Short Questions | 15 Long Questions | Solved Exercise

## Table of Contents

## Section 1 | Multiple Choice Questions

| # | Question | Options | Ans |
|---|---|---|---|
| 1 | What is the primary goal of data analytics? | (a) To store data<br>**(b) To find patterns and support decisions**<br>(c) To design websites<br>(d) To create games | B |
| 2 | Which of the following is a basic statistical concept? | **(a) Mean**<br>(b) Font size<br>(c) Pixel<br>(d) Bandwidth | A |
| 3 | What does the median represent? | (a) Most frequent value<br>**(b) Middle value in ordered data**<br>(c) Highest value<br>(d) Sum of values | B |
| 4 | What is the mode of a dataset? | (a) Average value<br>(b) Middle value<br>**(c) Most frequently occurring value**<br>(d) Smallest value | C |
| 5 | Which measure shows the spread of data? | (a) Mean<br>(b) Mode<br>(c) Median<br>**(d) Variance** | D |
| 6 | What does standard deviation measure? | (a) Central value<br>**(b) Data spread from mean**<br>(c) Data collection method<br>(d) Data storage | B |
| 7 | What is probability used for? | (a) Storing data<br>**(b) Predicting how likely an event is**<br>(c) Cleaning data<br>(d) Visualizing data | B |
| 8 | What is a common method of data collection? | **(a) Surveys**<br>(b) Formatting<br>(c) Compiling<br>(d) Encrypting | A |
| 9 | What does data cleaning involve? | (a) Adding new data<br>**(b) Fixing errors and inconsistencies**<br>(c) Designing charts<br>(d) Predicting outcomes | B |
| 10 | Why is data preparation important? | (a) It makes data colorful<br>**(b) It ensures accurate analysis**<br>(c) It deletes all data<br>(d) It slows down analysis | B |
| 11 | What is imputation? | (a) Removing all data<br>**(b) Replacing missing values with estimates**<br>(c) Sorting data<br>(d) Encrypting data | B |
| 12 | What is a statistical model used for? | (a) Designing logos<br>**(b) Analyzing data and making predictions** | B |

| | | | |
|---|---|---|---|
| | | (c) Formatting text<br>(d) Storing images | |
| 13 | What are the steps in building a statistical model? | **(a) Define, collect, choose, train, evaluate**<br>(b) Design, print, delete<br>(c) Format, encrypt, share<br>(d) Draw, paint, save | A |
| 14 | What is linear regression used for? | **(a) Predicting numerical outcomes**<br>(b) Grouping data<br>(c) Designing charts<br>(d) Storing files | A |
| 15 | What does the slope in linear regression represent? | (a) Starting value<br>**(b) Change in Y for a unit change in X**<br>(c) Total values<br>(d) Data spread | B |
| 16 | What is the intercept in linear regression? | (a) Change in Y<br>**(b) Value of Y when X is 0**<br>(c) Total data points<br>(d) Standard deviation | B |
| 17 | What does logistic regression predict? | (a) Continuous values<br>**(b) Binary or categorical outcomes**<br>(c) Data spread<br>(d) Missing values | B |
| 18 | What is the output range of logistic regression? | (a) -1 to 1<br>**(b) 0 to 1**<br>(c) 1 to 100<br>(d) Any real number | B |
| 19 | What is model evaluation? | (a) Designing a model<br>**(b) Testing how well a model performs**<br>(c) Deleting a model<br>(d) Formatting a model | B |
| 20 | What is an error metric used for? | **(a) Measuring prediction difference from actual values**<br>(b) Designing charts<br>(c) Collecting data<br>(d) Cleaning data | A |
| 21 | What does fairness and bias relate to in modeling? | (a) Data storage<br>**(b) Ethical considerations in modeling**<br>(c) Data visualization<br>(d) Data collection | B |
| 22 | What is data privacy concerned with? | **(a) Protecting personal information**<br>(b) Designing charts<br>(c) Cleaning data<br>(d) Collecting surveys | A |
| 23 | What is clustering used for? | **(a) Grouping similar items**<br>(b) Predicting numbers<br>(c) Testing models<br>(d) Cleaning data | A |
| 24 | What is K-means clustering? | (a) A regression model<br>**(b) A technique to group data into K clusters**<br>(c) A visualization tool<br>(d) A data cleaning method | B |

| | | | |
|---|---|---|---|
| 25 | What must be decided before running K-means clustering? | **(a) Number of clusters (K)**<br>(b) Type of chart<br>(c) Data source<br>(d) Regression formula | **A** |
| 26 | What is a bar chart best used for? | **(a) Comparing categories**<br>(b) Showing distribution<br>(c) Showing trends over time<br>(d) Showing relationships | **A** |
| 27 | What does a line graph show? | (a) Categories<br>**(b) Trends over time**<br>(c) Distribution<br>(d) Relationships | **B** |
| 28 | What does a histogram show? | (a) Trends over time<br>**(b) Distribution of data**<br>(c) Categories<br>(d) Relationships between two variables | **B** |
| 29 | What does a scatterplot show? | (a) Distribution<br>(b) Trends<br>**(c) Relationship between two variables**<br>(d) Categories | **C** |
| 30 | What does a boxplot display? | (a) Trends over time<br>**(b) Median, quartiles and outliers**<br>(c) Categories<br>(d) Bar heights | **B** |
| 31 | Which tool is commonly used for creating visualizations? | (a) MS Word<br>**(b) Excel**<br>(c) Notepad<br>(d) Paint | **B** |
| 32 | What is the square root of variance called? | (a) Mean<br>**(b) Standard deviation**<br>(c) Mode<br>(d) Median | **B** |
| 33 | What comes after model training in machine learning? | (a) Model deployment<br>**(b) Model evaluation**<br>(c) Data collection<br>(d) Data cleaning | **B** |
| 34 | What does β0 represent in linear regression? | (a) Slope<br>**(b) Intercept**<br>(c) Error term<br>(d) Dependent variable | **B** |
| 35 | What is the output of logistic regression? | (a) Continuous value<br>**(b) Binary outcome (0 or 1)**<br>(c) Time series<br>(d) Cluster | **B** |
| 36 | Which model predicts continuous numeric values? | (a) Logistic regression<br>**(b) Linear regression**<br>(c) K-means clustering<br>(d) Decision tree | **B** |
| 37 | Which loss function is used in linear regression? | (a) Log loss<br>(b) Cross-entropy loss | **C** |

| | | | |
|---|---|---|---|
| | | **(c) Mean squared error**<br>(d) Accuracy | |
| **38** | Which model uses a decision threshold? | (a) Linear regression<br>**(b) Logistic regression**<br>(c) K-means clustering<br>(d) PCA | **B** |
| **39** | What is the output range of logistic regression? | (a) -1 to 1<br>**(b) 0 to 1**<br>(c) Any real number<br>(d) 1 to 100 | **B** |
| **40** | What is clustering used for? | (a) Making predictions<br>**(b) Grouping similar items**<br>(c) Calculating probabilities<br>(d) Finding averages | **B** |
| **41** | Which clustering method requires choosing the number of groups beforehand? | (a) Hierarchical clustering<br>(b) DBSCAN<br>**(c) K-means clustering**<br>(d) Gaussian mixture model | **C** |
| **Introduction to Data Visualization** | | | |
| **42** | Which visualization is best for comparing categories? | (a) Line graph<br>(b) Histogram<br>**(c) Bar chart**<br>(d) Scatterplot | **C** |
| **43** | Which chart is used to show distribution of data? | (a) Line graph<br>(b) Pie chart<br>**(c) Histogram**<br>(d) Boxplot | **C** |
| **44** | Which tool is commonly used for creating visualizations? | (a) Word<br>**(b) Excel**<br>(c) Notepad<br>(d) Paint | **B** |
| **45** | What is the square root of variance? | (a) Mean<br>**(b) Standard deviation**<br>(c) Mode<br>(d) Median | **B** |
| **46** | What is the purpose of data transformation? | (a) To remove data<br>**(b) To make data suitable for analysis**<br>(c) To store data<br>(d) To hide data | **B** |
| **47** | Which of the following is not a type of data visualization? | (a) Bar chart<br>(b) Pie chart<br>**(c) Equation**<br>(d) Scatterplot | **C** |
| **48** | Which model predicts categorical outcomes? | (a) Linear regression<br>**(b) Logistic regression**<br>(c) PCA<br>(d) ARIMA | **B** |
| **49** | What does ethics in data science include? | (a) Data theft<br>**(b) Fairness and bias** | **B** |

| | | | |
|---|---|---|---|
| | | (c) Ignoring privacy<br>(d) Manipulating data | |
| 50 | What is the main goal of data analytics? | (a) Volume<br>**(b) Insights**<br>(c) Delete<br>(d) Backup | B |

## Answer Key

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

## Section 2 | Short Questions

| # | Question | Answer |
|---|---|---|
| | **Basic Statistical Concepts** | |
| 1 | **What is Data Analytics?** | Data analytics is the process of examining data to find patterns, draw conclusions and support decision-making. It helps turn raw data into useful insights. |
| 2 | **Why is data analytics important in business?** | It supports informed decisions and improves performance using data insights. It helps businesses identify opportunities and solve problems effectively. |
| 3 | **What are measures of central tendency?** | Mean, median and mode represent the center of a dataset. They help summarize large datasets with a single value. |
| 4 | **Define mean.** | Mean is the average of all numbers in a dataset. It is calculated by dividing the sum of values by the total number of values. |
| 5 | **How do you calculate the median?** | Arrange data in order and pick the middle value. If there's an even count, take the average of the two middle values. |
| 6 | **What is the mode in a dataset?** | Mode is the most frequently occurring value in a dataset. It can have more than one value if multiple numbers repeat equally. |
| 7 | **Can a dataset have more than one mode?** | Yes, if multiple values appear with equal highest frequency. This is known as bimodal or multimodal distribution. |
| | **Variance and Standard Deviation** | |
| 8 | **What is variance?** | Variance shows how spread out the data points are from the mean. A higher variance indicates greater variability in the data. |
| 9 | **What does high variance indicate?** | High variance means the data points are far from the mean. It suggests less consistency in the dataset. |
| 10 | **Define standard deviation.** | Standard deviation is the square root of variance. It tells us how much the data varies from the average. |
| 11 | **Why is standard deviation easier to interpret than variance?** | Because it's expressed in the same unit as the original data. Variance is in squared units, making it harder to interpret. |
| 12 | **What is the formula for standard deviation?** | √(Variance) or sqrt(sum of squared deviations divided by N). It gives a practical measure of data spread. |
| 13 | **What is probability?** | Probability measures how likely an event is to occur. It ranges from 0 (impossible) to 1 |

| | | |
|---|---|---|
| | | (certain). |
| 14 | **Give an example of probability in daily life.** | There's a 70% chance of rain tomorrow. This helps people decide whether to carry an umbrella. |
| 15 | **What is the formula for calculating probability?** | Probability = Favorable outcomes / Total possible outcomes. It helps quantify uncertainty in real-life situations. |
| 16 | **What is the probability of getting heads when flipping a coin?** | The probability is 1/2 or 50%. There are only two equally likely outcomes: heads or tails. |
| 17 | **How is probability used in weather forecasting?** | Meteorologists use probability to predict the likelihood of weather events. For example, 80% chance of thunderstorms. |
| **Data Collection and Preparation** | | |
| 18 | **What is data collection?** | It is gathering information from various sources for analysis. Accurate data collection ensures reliable results. |
| 19 | **Name three methods of collecting data.** | Surveys, observations and experiments. Each method has its own strengths and suitable applications. |
| 20 | **When would you use observation as a method?** | To study behavior without asking people directly. Example: Tracking customer movement in a store. |
| 21 | **What is the purpose of surveys?** | To gather opinions, preferences, or feedback from people. Surveys are easy to distribute and analyze. |
| 22 | **What is an experiment in data collection?** | An experiment tests cause-and-effect relationships. One variable is changed to see its effect on another. |
| 23 | **What is data preparation?** | It involves organizing and converting raw data into usable formats. Proper preparation ensures accurate analysis. |
| 24 | **What is data cleaning?** | It fixes or removes errors and inconsistencies in data. Clean data leads to better insights and decisions. |
| 25 | **Why is data cleaning important?** | Clean data ensures accurate and reliable analysis results. Unclean data can lead to misleading conclusions. |
| 26 | **What are some common issues found in unclean data?** | Missing values, duplicates, incorrect formatting and typos. These issues must be addressed before analysis. |
| 27 | **What is imputation?** | Imputation replaces missing values with estimated ones. Common methods include using the mean, median, or similar data. |
| **Building Statistical Modeling** | | |

| | | |
|---|---|---|
| 28 | **What is statistical modeling?** | It uses math models to analyze data and make predictions. Models help uncover patterns and forecast future outcomes. |
| 29 | **List the steps involved in building a statistical model.** | Define problem, collect data, choose model, train, evaluate. These steps ensure a structured approach to modeling. |
| 30 | **What is linear regression used for?** | It predicts numerical outcomes based on input variables. Example: Predicting house prices based on size and location. |
| 31 | **What is logistic regression used for?** | It predicts binary outcomes like yes/no or pass/fail. Example: Whether a student will pass or fail based on study hours. |
| 32 | **What is clustering?** | Clustering groups similar items together based on their features. It helps identify natural groupings in the data. |
| 33 | **What is K-means clustering?** | K-means divides data into K groups based on similarity. It's a simple and widely used clustering algorithm. |
| 34 | **How many clusters must you specify before running K-means?** | You must choose the number of clusters (K) first. Choosing the right K affects the quality of results. |
| 35 | **Give an example of using clustering in education.** | Grouping students based on academic performance for targeted teaching. Teachers can tailor strategies to each group. |
| 36 | **Why is clustering useful in market research?** | It segments customers with similar behaviors for marketing. Businesses can personalize offers to each segment. |
| **Introduction to Data Visualization** | | |
| 37 | **What is the purpose of data visualization?** | To represent data visually and help identify patterns, trends and insights. |
| 38 | **When would you use a bar chart?** | To compare different categories by showing their values as bars. |
| 39 | **What does a histogram show?** | The distribution of data by grouping values into intervals or bins. |
| 40 | **How can Excel or Google Sheets help with data visualization?** | They allow users to enter data and create charts like bar charts or line graphs for easier interpretation. |

## Section 3 | Long Questions

### Q.1 Define Data Analytics.

Data analytics is the process of examining, cleaning, transforming and interpreting raw data to discover meaningful patterns, draw conclusions and support decision-making. It involves using statistical and computational tools to analyze datasets and generate insights that can guide business strategies, improve operations, or solve problems.

### Q.2 What are the basic statistical concepts used in data analysis? Explain the measures of central tendency with suitable examples.

Statistics is a branch of mathematics that helps us understand and analyze data. By using statistics, large sets of information can be summarized in a simple way, making it easier to draw conclusions.

#### Measures of Central Tendency

Measures of central tendency help us identify the “center” or typical value in a dataset. There are three main measures of central tendency: mean, median and mode. These measures give us a sense of the average or most common values of a dataset.

#### Mean

The mean is the average of all the numbers in a dataset. To calculate the mean, we add all the numbers together and then divide the sum by the total number of values.

$$Mean = \frac{Sum\ of\ all\ values}{Total\ number\ of\ values}$$

**Example:** Imagine 5 students scored 50, 60, 70, 80 and 90 in a test. The mean score is calculated by adding all the scores and then dividing by the number of students:

$$Mean = \frac{50 + 60 + 70 + 80 + 90}{5} = 70$$

#### Median

The median is the middle value in a dataset when the numbers are arranged in order. If there is an odd number of values, the median is the exact middle number. If there is an even number of values, the median is the average of the two middle numbers.

**Example:** Using the same test scores: 50, 60, 70, 80 and 90. When we arrange these scores in ascending order (which they already are), the middle value is 70. Therefore, the median score is 70.

**Example:** Example with Even Number of values: If the scores were 50, 60, 70 and 80, we would take the average of the two middle scores (60 and 70): Median = (60+70)/2 = 65, so the median is 65. The median helps us understand the middle point of the data.

#### Mode

The mode is the number that appears most often in a data set. There can be more than one mode if multiple numbers appear with the same highest frequency. The mode helps us identify the most frequent or common value in the data.

**Example:** If 5 students scored 50, 60, 70, 70 and 90, the number 70 appears twice, while all other numbers appear only once. Therefore, the mode is 70.

**Example:** Example with Multiple Modes: If the scores were 50, 60, 70, 70, 60 and 90, both 60 and 70 appear twice. So, there are two modes: 60 and 70.

## Q.3 Explain measures of dispersion with suitable example.

Measures of dispersion tell us how spread out or scattered the data is. Two common measures of dispersion are variance and standard deviation. These help us understand whether the data points are close to the average (mean) or spread far from it.

Two common ways to measure dispersion are: 1) Variance 2) Standard Deviation

### 1) Variance

The variance shows how much the numbers in a data set differ from the mean. A higher variance means that the numbers are more spread out, while a lower variance means that the numbers are closer to the mean. To calculate variance, we use the following mathematical formula.

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

Where: x<sub>i</sub> = represents each individual value in the data set. μ = is the mean of the data set. N = is the total number of values in the data set.

**Example:** For the following two classes find out which class is more spread out by calculating their variance? Class A: 50, 52, 55, 57, 60. Class B: 30, 45, 55, 75, 90.

### Step 1: Variance for Class A

Given Score = 50, 52, 55, 57, 60

$$\mu = \frac{50+52+55+57+60}{5} = \frac{274}{5} = 54.8$$

#### Step 1.2: Compute Each Squared Deviation $(x_i - \mu)^2$

| $x_i$ | $x_i - \mu$ | $(x_i - \mu)^2$ |
|---|---|---|
| 50 | 50 − 54.8 = −4.8 | 23.04 |
| 52 | 52 − 54.8 = −2.8 | 7.84 |
| 55 | 55 − 54.8 = 0.2 | 0.04 |
| 57 | 57 − 54.8 = 2.2 | 4.84 |
| 60 | 60 − 54.8 = 5.2 | 27.04 |

### Step 1.3: Compute Variance

$$Variance\ (\sigma^2) = \frac{23.04 + 7.84 + 0.04 + 4.84 + 27.04}{5}$$

$$\sigma^2 = 62.8 / 5 = 12.56$$

### Step 2: Variance for Class B

Given Scores: 30, 45, 55, 75, 90

$$\mu = \frac{30+45+55+75+90}{5} = \frac{295}{5} = 59$$

| $x_i$ | $x_i - \mu$ | $(x_i - \mu)^2$ |
|---|---|---|
| 30 | 30 − 59 = −29 | 841 |
| 45 | 45 − 59 = −14 | 196 |
| 55 | 55 − 59 = −4 | 16 |
| 75 | 75 − 59 = 16 | 256 |
| 90 | 90 − 59 = 31 | 961 |

### Step 3: Compute Variance

$$\sigma^2 = \frac{841 + 196 + 16 + 256 + 961}{5} = 2270 / 5 = 454$$

- Variance of Class A: 12.56
- Variance of Class B: 454

This confirms that Class B has a much higher variance, meaning the scores are more spread out compared to Class A.

### 2) Standard Deviation

It is similar to variance but provides a more practical and interpretable value because it is in the same unit as the original data. The standard deviation tells us how spread out the numbers are in relation to the mean. The standard deviation is simply the square root of the variance. To calculate standard deviation, we use the following mathematical formula.

$$\text{Standard Deviation} = \sigma = \sqrt{\left[\frac{1}{N}\ \Sigma\ (x_i - \mu)^2\right]}$$

Where: x represents each individual value in the data set. μ is the mean of the data set. N is the total number of values in the data set.

**Example:** Calculating Standard Deviation:

Class A: Standard Deviation = √12.56 = 3.55

Class B: Standard Deviation = √454 = 21.31

The standard deviation for Class A is approximately 3.55, while for Class B, it is about 21.31. This means that Class A's scores are closely packed around the mean, whereas Class B's scores are more widely scattered. The standard deviation helps us easily understand how much variation there is in the scores.

**CLASS ACTIVITY**

**Instructions:** You will analyze a small dataset, calculate measures of central tendency (mean, median, mode) and measures of dispersion (variance and standard deviation).

- Collect Data: Imagine you are surveying your classmates about the number of hours they spend on homework in a week. Gather data from 10 classmates. Record the number of hours (you can use any reasonable numbers, e.g., between 0 to 20 hours).
- Calculate Measures of Central Tendency — Mean, Median, Mode.
- Calculate Measures of Dispersion — Variance, Standard Deviation.
- Reflect: Write a brief reflection (3-4 sentences) about what these calculations tell you regarding your classmates' study habits.

**Solution**

**1) Data Collection:** Students surveyed 10 classmates and recorded how many hours they spend on homework weekly.
**Collected Data:** 5, 7, 6, 8, 7, 10, 7, 9, 6, 8

**2) Measures of Central Tendency**

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}} = \frac{73}{10} = 7.3$$

On average, students spend 7.3 hours per week on homework.

Ordered data: 5, 6, 6, 7, 7, 7, 8, 8, 9, 10
Since there are 10 numbers (even), the median is the average of the two middle numbers:

$$\text{Median} = \frac{7 + 7}{2} = 7$$

The middle value is 7 hours.

**Mode:** The number 7 appears three times — more than any other number. The most common study time is 7 hours.

**3) Measures of Dispersion**

$$\sigma^2 = \frac{1}{N}\Sigma(x_i - \mu)^2 = \frac{20.10}{10} = 2.01$$

A variance of 2.01 shows the data points are fairly close to the mean.

$$\sigma = \sqrt{2.01} \approx 1.42$$

On average, students' homework times differ from the mean by about 1.42 hours.

**4) Reflection**

- Most students spend around 7 hours on homework each week.
- There's not much variation in study habits — most students are close to the average.
- The mode and median being the same reinforces that 7 hours is a typical amount of study time.
- The low standard deviation confirms that individual differences are small.

**DO YOU KNOW?**

**Q.** Where is statistics used in everyday life?
**Ans:** Statistics is used in many everyday activities, such as predicting weather patterns, analyzing sports performance, making business decisions and understanding trends in social media or health data.

## Q.4 What is Probability? Explain with example.

Probability is the study of how likely an event is to happen. It helps us make predictions based on known information.

**Example:** Consider flipping a coin. There are two possible outcomes: heads or tails. Since both outcomes are equally likely, the probability of getting heads is 50% (or 1/2) and the probability of getting tails is also 50%.

We can express this mathematically as:

$$Probability = \frac{Number\ of\ favorable\ outcomes}{Total\ number\ of\ outcomes}$$

### In the case of the coin flip

Probability of heads = (1 favorable outcome, 2 total outcomes) = 1/2

Probability is not just for coin flips. It is used in many areas, such as predicting the weather, making business decisions, or even predicting outcomes in sports like cricket.

### Real-Life Uses of Probability

Probability isn't just for games or coins — it's used in many real-life situations, like:

- Predicting weather forecasts (e.g., 70% chance of rain)
- Making business decisions (e.g., risk analysis)
- Playing games like cricket, where players guess the chance of scoring runs
- Analyzing data in research or school projects

**TIDBITS**

**Q.** How can statistics help in decision-making?
**Ans:** Statistics helps us understand patterns in data, leading to better and more informed decisions in fields like healthcare, marketing, education and more.

## Q.5 Explain data collection methods.

Data collection refers to the process of gathering relevant information for a particular purpose. Depending on the nature of the research, different methods can be used for data collection. These methods include surveys, observations and experiments, each having its own strengths and appropriate contexts. Choosing the right method depends on the research objective and the type of data required.

### 1) Surveys

Surveys are a commonly used method for collecting large amounts of data in a structured way. They involve asking a predefined set of questions to a sample group. Surveys can be conducted using various means such as online forms, telephone calls, or face-to-face interviews.

**Example:** A small local grocery store in Islamabad wants to know customer preferences regarding which products they would like to see more frequently. The store creates a short survey consisting of five questions as shown below and distributes it to 50 customers over the weekend. The collected responses are then analyzed to stock products that align with customer demand, helping improve business operations.

Customer Preference Survey/Questions:

- Which product categories do you buy most often? (e.g., fruits, vegetables, dairy)
- Are there any products you would like to see more often?
- How often do you shop at this grocery store? (e.g., daily, weekly, monthly)
- What influences your purchasing decisions the most? (e.g., price, quality, availability)
- Any additional comments or suggestions?

### 2) Observations

Observation involves collecting data by watching or monitoring subjects in their natural environment. This method is useful when researchers want to gather data on behaviors or phenomena without interference.

**Example:** A restaurant is interested in knowing which tables are most frequently chosen by customers during lunchtime. A staff member observes the seating choices over a period of one week. Based on this observation, the restaurant arranges its seating to optimize the customer comfort and traffic flow, which helps in improving customer satisfaction and service efficiency.

### 3) Experiments

Experiments involve manipulating one or more variables to determine their effect on another variable. This method is particularly useful in scientific and engineering fields where controlled environments are necessary for accurate measurement.

**Example:** A school teacher wants to test whether providing students with printed notes helps improve their performance in exams. The teacher conducts an experiment with two groups of students, one receiving printed notes and the other relying solely on lectures. After one month, both groups took the same test and the teacher compared the results to see if printed notes had a positive impact on academic performance.

## Q.6 What is Data Preparation? Explain how missing or incorrect data is handled. Support your answer with examples.

Once data has been collected, it is important to prepare it for analysis. This includes cleaning the data to remove errors or inconsistencies, organizing it in a meaningful way and converting it into a format suitable for analysis. In cases where data is missing or incorrect, researchers may need to employ techniques such as interpolation or statistical adjustments to ensure the accuracy and reliability of the results.

**Example:** If survey responses contain incomplete information, missing values can be estimated based on the available data. Proper data preparation ensures that the analysis leads to reliable and valid results.

### Data Cleaning

Data cleaning means correcting or removing any problems in the data. These problems can include incorrect entries, missing values, or duplicate results. If these errors are not fixed, the results of the analysis will be misleading.

**Example:** Imagine a school collecting data on student scores. Some students may have entered their names incorrectly, or a few scores may be missing from the records. In this case, data cleaning would involve correcting any wrong names and finding the missing grades to complete the dataset.

Following table, illustrates the data cleaning process for student scores in a school. It shows examples of common issues such as incorrect names, missing scores and duplicate entries.

**Original Data (with Errors)**

| Name | Scores | Class | Section |
|---|---|---|---|
| Ali | 84 | 10 | A |
| Alie | 90 | 10 | A |
| Sara | — (missing) | 10 | A |

**Cleaned Data (After Data Cleaning)**

| Name | Grade | Class | Section |
|---|---|---|---|
| Ali | 84 | 10 | A |
| Ali | 90 | 10 | A |
| Sara | 87 | 10 | A |

### Data Transformation

Once the data is clean, it often needs to be transformed into a format that is easier to work with. This transformation may include converting data into different formats, creating new columns, or organizing data in a different way. These changes help make the data more suitable for analysis or modeling.

**Example:** After cleaning the student grade records, it may be necessary to transform this data for better analysis. For instance, instead of displaying grades for each individual student, the data might be aggregated or summarized to show class-level statistics such as average scores or grade distribution.

### Handling Missing Data

Sometimes, data is incomplete or has missing values. There are different techniques to handle missing data. One option is to remove the rows with missing values if they are very few. Another option is to fill in the missing values with an average or with data from similar cases. The choice depends on the type of data and the amount of missing information.

**Example:** In the dataset of student grades, if Sara's grade is missing, this creates a challenge in assessing her performance. To address this issue, several strategies can be employed:

- **Imputation:** One common method is to estimate the missing value using existing data. Example: The school can calculate the average score of all students in Sara's class. If the average score is 87, the school may assign this value to Sara's record temporarily. This approach allows the school to maintain a complete dataset while making a reasonable assumption about Sara's performance.

- **Flagging:** The school can also keep track of Sara's missing score by adding a note in the dataset. This method indicates that Sara's score is not available, making analysts aware of the incomplete data. This approach ensures transparency while allowing the analysis to proceed without filling in the gap.
- **Removal:** If the number of missing entries is small, the school might choose to exclude Sara's record from specific analyses. This decision is acceptable if it does not significantly impact the overall understanding of student performance. However, it risks losing valuable information about Sara.

## Q.7 What is Statistical modeling and how to build a statistical model?

Statistical modeling is an approach used to analyse data to make sense of the real-world and to predict what will happen in the future. Think of it like this: if you want to know how much money you'll spend on groceries next month, you can look at your past expenses. By analyzing that data, you can create a model to help you estimate your future grocery expenses.

### Model Development

Building a statistical model involves several steps:

- **Step 1: Define the Problem:** First, we need to understand the problem. Example: If we are trying to predict grocery expenses, we need to identify the factors that influence them (e.g., family size, location, or income).
- **Step 2: Collect Data:** Next, we gather data related to the problem. Example: We will collect data on past spending habits, number of family members and any other factors that may affect grocery costs.
- **Step 3: Choose an Algorithm:** Based on the problem and the data, we choose an appropriate algorithm. Algorithms are methods that help us build models. Some popular algorithms are linear regression and logistic regression.
- **Step 4: Train the Model:** The model is then trained using the collected data. This means the model learns from the data to make predictions.
- **Step 5: Evaluate the Model:** Finally, we test the model to see how well it works by using new or unseen data. This step is very important to ensure the model makes accurate predictions.

## Q.8 Explain Linear Regression with example.

Linear regression is a widely used statistical model that helps understand the relationship between two variables. It is often used to predict one variable based on another.

**Example:** Imagine you run a small fruit stall in your town and you want to predict how much money you will make each day based on the number of customers who visit your stall. The number of customers is the independent variable (the cause) and the money you earn is the dependent variable (the effect). We will use linear regression to understand this relationship and help you predict future earnings.

### Step 1: Collecting Data

To build a linear regression model, we need data. Let's assume you've recorded the number of customers and your daily earnings for the last 5 days:

**Customer's Data**

| Number of Customers | Daily Earnings (Rs.) |
|---|---|
| 10 | 500 |
| 15 | 700 |
| 20 | 900 |
| 25 | 1,100 |
| 30 | 1,300 |

Here, the number of customers is our independent variable (X) and the daily earnings are the dependent variable (Y).

### Step 2: Understanding the Linear Regression Formula

The formula for simple linear regression is:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Where: Y is the dependent variable (in our case, daily earnings), X is the independent variable (the number of customers), β<sub>0</sub> is the intercept, which is the starting value of Y when X = 0, β<sub>1</sub> is the slope, which shows how much Y changes with each unit increase in X, ε is the error term, which accounts for the difference between the predicted and actual values.

### Step 3: Building the Linear Regression Model

When building a linear regression model, our goal is to find the best line that explains how two things are related — in this case, the number of customers and daily earnings. Here's how we get the values for the slope (40) and intercept (100):

### Understanding the Slope ($\beta_1 = 40$)

The slope shows how much extra money we make for every new customer. Let's use our data to figure it out: If you notice, for every 5 extra customers, earnings go up by 200 rupees. So, for each new customer:

$$\beta_1 = 200 / 5 = 40$$

This means every new customer adds 40 rupees to our earnings.

### Understanding the Intercept ($\beta_0 = 100$)

The intercept ($\beta_0$) represents the earnings when no customers visit. To find this value, we look at where the line crosses the vertical axis when the number of customers is zero. In simpler terms, it tells us what the base earnings are, even if no one shows up.

We can use the equation: Earnings = $\beta_0 + \beta_1 \times$ Customers. If we take any data point, say when there are 10 customers, the earnings are 500 rupees. Substituting these values into the equation:

$$500 = \beta_0 + (40 \times 10)$$

$$500 = \beta_0 + 400$$

### Solving this gives

$$\beta_0 = 500 - 400 = 100$$

This means that, based on the data, if no customers show up, you'd still expect to make 100 rupees, maybe from regular customers or other fixed earnings. So, the intercept value of 100 rupees represents the minimum amount you'd make on a day with zero customers.

### Final Equation

$$Earnings = 100 + 40 \times Customers$$

This equation means: You'll always earn 100 rupees, even if no one comes. Each new customer adds 40 Rs to your total.

### Step 4: Interpreting the Model

Once the model is built, we can use it to predict future earnings.

**Example:** If you expect 22 customers tomorrow, the predicted earnings would be:

$$Earnings = 100 + (40 \times 22) = 100 + 880 = 980\ Rs$$

This means that with 22 customers, you can expect to earn around 980 rupees.

### Step 5: Testing the Model

After building the model, it's important to test it using new data. Let's say on the 6th day, 28 customers visited your stall and you earned 1,250 Rs. Using the model, we can predict the earnings for 28 customers:

$$Predicted\ Earnings = 100 + (40 \times 28) = 100 + 1{,}120 = 1{,}220\ Rs$$

However, you actually earned 1,250 Rs. The difference between the predicted and actual earnings is called the error:

$$Error = 1{,}220 - 1{,}250 = -30\ Rs$$

While the prediction was close, it is not perfect, showing that real-world data often has some variation.

> **TIDBITS**
>
> **Q.** How can you make your statistical model more accurate?
> **Ans:** Use more data points to help the model learn better.
>
> **Q.** What should you include in your model to improve its predictions?
> **Ans:** Include important factors like family size or special events that affect spending.
>
> **Q.** Why is it important to update your model with new data?
> **Ans:** To keep the model up-to-date and improve its predictions over time.
>
> **Q.** How can you check if your model is working well?
> **Ans:** Compare your model's predictions with actual results to see where it can be improved.

## Q.9 Define Logistic Regression.

Logistic regression is a powerful tool used when we want to predict an outcome that can be categorized as “yes” or “no”.

**Example:** Let's say we want to determine whether a student will pass or fail an exam based on the number of hours they study. Instead of predicting a specific score, logistic regression helps us find the probability of passing.

### Understanding Logistic Regression

Logistic regression is different from linear regression because it does not predict exact numbers. Instead, it provides a probability value between 0 and 1. This means it tells us how likely something is to happen.

## Q.10 What is the difference between Logistic regression and Linear regression?

Difference between Logistic regression and Linear regression:

| Feature | Linear Regression | Logistic Regression |
|---|---|---|
| Purpose | Predicts continuous numeric values | Predicts categorical outcomes (usually binary: 0 or 1) |
| Output | Real number (e.g., price, age, temperature) | Probability (between 0 and 1), then mapped to classes |
| Algorithm Type | Regression | Classification |
| Equation | $y = \beta_0 + \beta_1 x + \ldots + \beta_n x_n$ | $p = 1 / (1 + e^{-(\beta_0+\beta_1 x+\ldots+\beta_n x_n)})$ |
| Output Range | $-\infty$ to $+\infty$ | 0 to 1 |
| Loss Function | Mean Squared Error (MSE) | Log Loss (Cross-Entropy Loss) |
| Use Case Examples | Predicting house prices, sales, etc. | Email spam detection, disease prediction, etc. |
| Decision Boundary | Not required | Yes, decision threshold (commonly 0.5) is used |

## Q.11 Explain Clustering Techniques.

Clustering is a way of grouping similar things together based on their characteristics. Imagine you have a group of students in your class and you want to divide them into groups based on their performance in different subjects like math, science and English. Clustering helps us do that by creating groups of students who perform similarly.

**Example:** Clustering of Students by Performance. Let's say we have data for five students, showing their scores in Math and English:

**Data for Clustering Techniques**

| Student | Math Score | English Score |
|---|---|---|
| Arshad | 85 | 70 |
| Ayaan | 90 | 65 |
| Areeda | 50 | 80 |
| Sumera | 40 | 85 |
| Wayna | 60 | 60 |

We can use clustering to group these students based on their performance in these two subjects.

## Q.12 What is K-means clustering?

K-means clustering is one of the simplest and most popular techniques to group data. In K-means, we need to decide how many groups (clusters) we want.

**Example:** Let's say we want to divide the students into two clusters: one for students who are strong in math and one for students who are strong in English.

The algorithm will group students with similar performance in Math and English together by calculating the distance between their scores and finding patterns. It will assign students like Arshad and Ayaan (who are good at math) to one group and students like Areeda and Sumera (who are good at English) to another group.

## Q.13 How do we evaluate models? Also explain the importance of ethical considerations while using data.

Evaluating and Understanding Models. Once a model is built, it is important to check how well it performs and to understand the results it provides. This is called model evaluation.

### 1) Performance Metrics

Performance metrics help us measure how well a model is doing. Some common metrics include:

- **Error Metrics:** Error metrics measure how much the model's predictions differ from the actual values. In our grocery example, if the model predicts a monthly grocery bill of 8,000 rupees but the actual bill is 10,000 rupees, the difference is the error.
- **Accuracy Metrics:** Accuracy metrics tell us how many of the model's predictions were correct. For example, if a model predicts whether a student will pass or fail an exam, accuracy measures the percentage of correct predictions made by the model.

### 2) Interpreting Outputs

Interpreting a model's output means understanding what the results reveal.

- **Drawing Conclusions from Insights:** For example, if our linear regression model shows that number of hours studied strongly affects exam scores, we can conclude that students should study more hours to improve their scores.

### Ethical Considerations in Modeling

When building models, it is important to consider the ethical implications, such as fairness and privacy.

- **Fairness and Bias:** A model should be fair and unbiased. Example: If a model is used to decide who gets a loan, it should not unfairly favor one group of people over another.
- **Data Privacy:** We must protect personal information used in models. Example: A company using customer data should keep it safe and not share it without permission.

> **TIDBITS**
>
> **Q.** Why is it important to visualize your data before building models and to test your results on new data?
> **Ans:** Visualizing your data helps you understand patterns, spot outliers or errors and choose the right model. Testing your results on new data ensures your model works well in real-life situations and gives accurate predictions. Both steps help build better and more reliable models.

## Q.14 Discuss the types of data visualizations and their uses.

Data visualization is the process of representing data in a visual format, such as graphs or charts. It helps us to quickly identify patterns, trends and insights from the data.

### Types of Visualizations

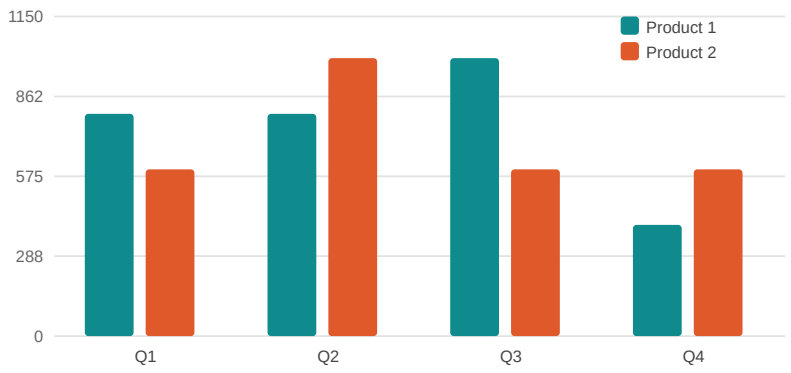
Data visualization is a powerful way to understand complex information. Different types of visualizations serve various purposes, making it easier to interpret and analyze data.

### 1) Bar Charts

Bar charts are ideal for comparing different categories. Each bar represents a category and the height (or length) of the bar indicates the value associated with that category.

Use: Bar charts help us compare different things, like how many students are in each section or which product sold the most.

**Example:** Imagine you want to compare the sales figures for different products in a store. A bar chart can visually represent this data.



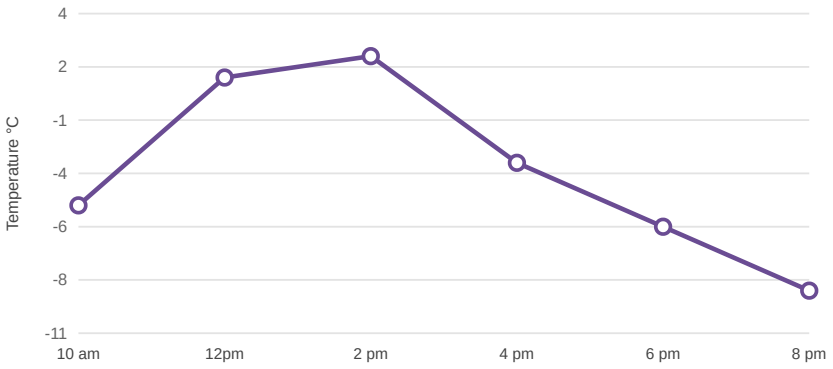

A bar chart showing sales figures of different products

### 2) Line Graphs

Line graphs are used to show trends over time. They plot data points and connect them with a line, making it easy to observe changes.

Uses: Line charts are good for showing changes over time, like how the temperature changes each day or how expenses increase each month.

**Example:** If you track the temperature over a week, a line graph will show how the temperature rises and falls each day.



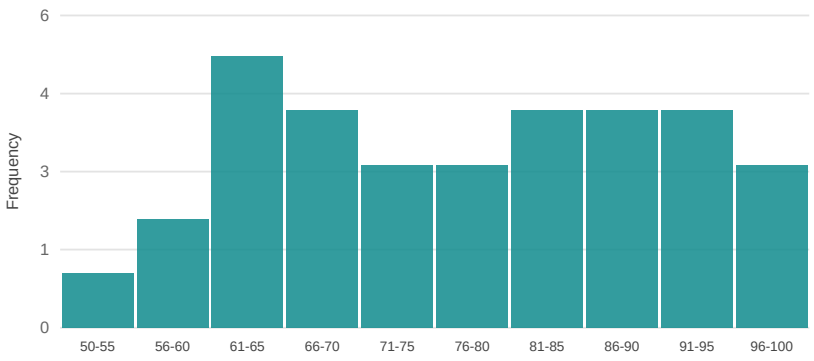

A line graph showing variation of temperature over time

### 3) Histograms

Histograms are used to show the distribution of a dataset. They group data into bins or intervals, allowing you to see how frequently values occur within those ranges.

Use: Histograms show how data is spread out, like marks in a test or heights of students.

**Example:** If you want to analyze how students performed in a math exam, a histogram can show the distribution of scores.



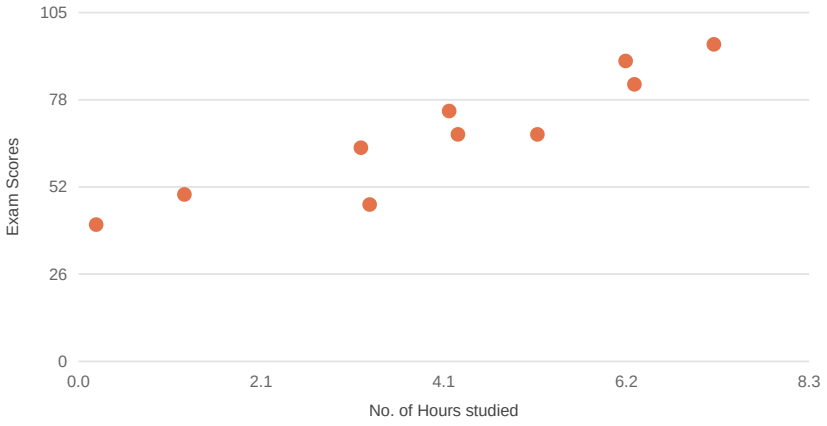

**Example of a Histogram showing the distribution of exam scores**

### 4) Scatterplots

Scatterplots display relationships between two variables. Each point represents an observation and the position on the graph indicates values for both variables.

Use: Scatterplots help us see if two things are connected, like if more study time leads to better marks.

**Example:** A scatterplot can be used to explore the relationship between the number of hours studied and exam scores.



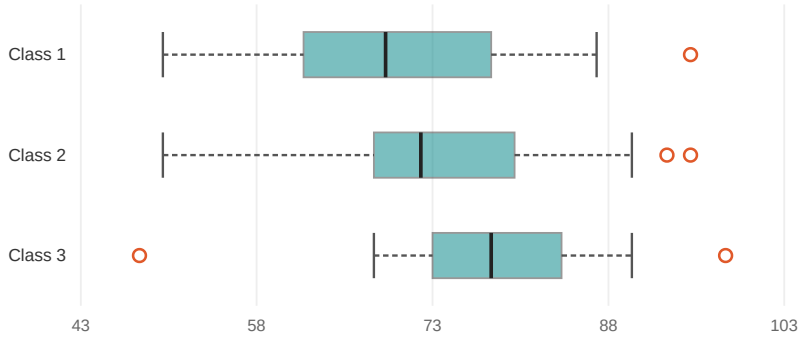

**Scatterplot showing the relationship between hours studied and exam scores**

### 5) Boxplots

Boxplots, or whisker plots, summarize data distribution by displaying the median, quartiles and potential outliers. They provide a visual summary of data variability.

Use: Boxplots are useful to compare groups, like marks of students in different classes.

**Example:** A boxplot can be used to compare the exam scores of different classes to see which class performed better overall.




A Boxplot, showing class scores performance of three classes

## Q.15 Explain the tools used for data visualization and describe the steps to create and interpret visualization using Excel and Google sheets.

Data visualization means showing data in charts, graphs, or maps so it's easier to understand.

### Using Excel and Google Sheets for Visualization

Excel and Google Sheets: These tools allow you to easily enter your data and then generate a variety of visualizations such as bar charts, line graphs and pie charts.

**Example:** Let's say you run a small business in your local area and you want to track how many products you sell each month. You can enter the data for each month in Excel or Google Sheets and with just a few clicks, you can create a bar chart to see which month had the most sales.

### Creating and Interpreting Visualizations

Step-by-Step Guide: Here's a simple guide to creating a visualization in Excel or Google Sheets.

- **Enter Your Data:** Start by entering your data into the spreadsheet. Example: In one column, you could have the months (January, February, etc.) and in another column, the sales figures for each month.
- **Select the Data:** Highlight the data you want to visualize by clicking and dragging your mouse over the cells.
- **Choose a Chart Type:** Click on the "Insert" tab and select the type of chart you want to create (bar chart, line graph, etc.).
- **Customize the Chart:** You can add labels to your chart to make it clearer, such as labeling the x-axis with the months and the y-axis with the sales figures. This makes the chart easier to interpret.
- **Understanding Statistical Representations:** When you create a visualization, it's important to understand what the chart is telling you.

## Section 4 | Solved Textbook Exercise

### Multiple Choice Questions (MCQs)

| # | Question | Options | Ans |
|---|---|---|---|
| 1 | An example of a basic statistical model: | **(a) Linear Regression**<br>(b) Neural Networks<br>(c) Decision Trees<br>(d) Support Vector Machines | A |
| 2 | The activity involved in experimental design in data science: | (a) Creating visualizations<br>**(b) Collecting and analyzing data systematically**<br>(c) Writing code for machine learning<br>(d) Building databases | B |
| 3 | A commonly used tool for creating data visualizations: | (a) MS Excel<br>(b) Python (Matplotlib)<br>(c) Tableau<br>**(d) All of the above** | D |
| 4 | The meaning of the slope in a linear regression model: | (a) The intercept of the model<br>**(b) The change in the dependent variable for a unit change in the independent variable**<br>(c) The error term<br>(d) The mean of the data | B |
| 5 | An example of a real-world application of statistical models: | **(a) Predicting house prices**<br>(b) Creating social media posts<br>(c) Designing websites<br>(d) Writing essays | A |
| 6 | Option not considered a benefit of data visualization: | (a) Identifying trends and patterns<br>(b) Communicating insights effectively<br>**(c) Making data more complex**<br>(d) Summarizing large datasets | C |
| 7 | A primary goal of K-Means Clustering: | (a) To classify data into predefined categories<br>**(b) To group data into clusters based on similarity**<br>(c) To predict continuous outcomes<br>(d) To reduce the dimensionality of data | B |
| 8 | The meaning of "K" in K-Means Clustering: | (a) Number of features in the dataset<br>**(b) Number of clusters to be formed**<br>(c) Number of iterations required for convergence<br>(d) Number of data points in the dataset | B |

### Answer Key

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | A | 2 | B | 3 | D | 4 | B | 5 | A | 6 | C | 7 | B | 8 | B |

### Short Questions

| # | Question | Answer |
|---|---|---|

| | | |
|---|---|---|
| 1 | **What is the importance of building statistical models in real-world applications?** | Statistical models help us understand patterns, make predictions and support decision-making in areas like business, education, healthcare and more. |
| 2 | **Name one basic statistical model used for predicting outcomes and explain its purpose.** | Linear regression is a basic model used to predict a numerical outcome (like exam scores) based on one or more input variables (like study hours). |
| 3 | **List two types of data visualizations and describe when you would use each.** | Bar chart: Used to compare values across categories (e.g., sales per month). Line graph: Used to show trends over time (e.g., temperature changes during the year). |
| 4 | **How does visualizing data help in understanding descriptive statistics?** | Visualizing data makes it easier to see patterns, averages and variations in the data quickly, helping us understand key statistics like mean, median and spread without just looking at numbers. |

## Long Questions

### Q.1 Explain the role and importance of statistical models in solving real-world problems.

Statistical models help us understand patterns in data and make predictions about future events. They play a key role in solving real-world problems by turning raw data into useful information.

**Examples**

- A shop owner can use a model to predict how much stock will be needed next month.
- A school can predict student performance based on past results.
- Doctors can use models to understand how diseases spread.

These models help in making better decisions, saving time and improving outcomes in fields like business, education, healthcare and finance.

### Q.2 Describe the steps involved in building a basic statistical model (e.g., linear regression). Include details on data collection, model training and evaluation.

**Ans:** See Q.8 of the Long Questions section (Section 3) for the complete answer.

### Q.3 Discuss the types of data visualizations and their uses.

**Ans:** See Q.14 of the Long Questions section (Section 3) for the complete answer.

### Q.4 Explain data collection methods.

**Ans:** See Q.5 of the Long Questions section (Section 3) for the complete answer.

### Q.5 Discuss the concept of measure of tendency with example.

**Ans:** See Q.2 of the Long Questions section (Section 3) for the complete answer.