Class IX

# ANSWER KEY

## Data Science

**Part-B: Subject Specific Skills**

# 1. Introduction

## Unsolved Exercise

### Objective Type Questions

**A.** 1. a 2. c 3. b 4. c

**B.** 1. F 2. T 3. T 4. F

### Standard Questions

**A.** 1. Four qualities of information is as follows:

- Availability
- Accuracy
- Timeliness
- Completeness

2. The objective of IT is to promote and encourage interactions among professionals from practice and research and advancement of investigation of concepts, methods, techniques, tools and issues related to information systems in organisations.
3. Using Knowledge means the appropriate collection of information that can make it useful.
4. Data has influenced many areas of the healthcare industry.
   - Tracking the medical and health history of patients
   - Prediction of disease transmission and epidemics
   - Treatment of COVID-19 pandemic affected people
5. 
   - Friends and family might track you out of curiosity.
   - Companies might track you to build a profile on you.
   - Employers might track you to make human resources decisions.
   - Cybercriminals might track you for a variety of reasons.

**B.** 1. There are various industries that are rapidly influenced from data. Some of them are:

- **Healthcare:** Healthcare industry is the most vulnerable sector utilising modern available data.
- **Online Shopping:** Online seller's analyse customer's historical purchases, searches, etc. to come up with targeted marketing. They reach out to consumers with curated offers and advertisements.

- **Education:** In most of the universities, colleges and schools, the admission process is now digitised. Also, students explore various career options by analysing historical records of universities and educational institutes.
- **Online Shows:** Due to advancement in data analysis techniques, you can watch most online streaming platforms recommending your personalized content.

2. Data can be lost due to a system crash when the system stops abruptly. For example, issues in the power supply may cause hardware or software failure. Another reason for losing data can be a disk failure when the hard disk drives or storage drives fail. Examples include damage to the storage drive, formation of bad sectors, etc.
3. Your data footprint impact may, be larger than how much physical or online data storage you have. There are two types of digital footprints which are passive and active.

   An 'active' digital footprint is the publicly traceable information that you share on the web, including Facebook updates, message board posts and Twitter rants.

   Your passive digital footprint is made up of the information that companies are harvesting behind the scenes, such as browsing data, IP addresses and purchasing habits.
4. The DIKW Pyramid describes the relationships between data, information, knowledge and wisdom. Ultimately, it guides in achieving a particular goal of any organisation. Each building block is a step towards a higher level. First comes data, then information, next is knowledge and finally comes wisdom. Each step answers different questions about the initial data and adds value to it. The more we enrich our data with meaning and context, the more knowledge and insights we get out of it so we can take better, informed and data-based decisions.
5. The process of restoring the inaccessible, lost, corrupted, damaged or deleted data is called data recovery. There could be quite several reasons for data loss. To prevent this, we should frequently back up our data. Large enterprise systems generally use backup data storage from where they recover the data in case of any loss.
6. Due to advancement in data analysis techniques, you can watch most online streaming platforms recommending your personalized content. They record previous watch history and determine which actors, genres, concepts appeal more to the viewers. They also provide users with ratings and feedback based on the historical input of other viewers. They also predict which shows will become popular.
7. Your personal information like name, email address, contact number, address, etc. that you have shared on the Internet while registering on a website is known as your personal data. You can keep your personal data safe by ensuring the authenticity of the website on which you are registering.

## Higher Order Thinking Skills

Do yourself.

# 2. Arranging and Collecting Data

## Unsolved Exercise

### Objective Type Questions

**A.** 1. b 2. a 3. b 4. b 5. d 6. d
7. b 8. d 9. c 10. a 11. d 12. d
13. a

**B.** 1. F 2. T 3. T 4. F 5. F

### Standard Questions

**A.** 1. As data scientist, you need to collect data for various purposes. For generating insights for business purpose, your attention is more on providing means of achieving business goals mostly for profits making.

A researcher or scientist works based on the collected data. Data collection is a primary and essential step in most cases.

2. Improper data collection leads to:
   - Inability to answer research questions accurately
   - Inability to repeat and validate the study
   - Distorted findings resulting in wasted resources
   - Misleading other researchers to pursue fruitless avenues of investigation
   - Compromising decisions for public policy
   - Causing harm to human participants and animal subjects

3. Data collection tool does the following:
   - Create: Build smart surveys with the right questions to gather meaningful data
   - Collect: Gather data by employing various sharing-channels, online and offline
   - Study: Analyse the survey data to uncover valuable insights from the results
   - Act: Take actionable measures based on the rich survey insights and improve

4. This would be ordinal variable.
5. Big Data does not have to be big (in Petabytes/Exabytes). Even 50 GB can be said as big data if the structure is too complex for a normal Relational Database Management System (RDBMS) to store.
6. The 5 V's of big data are: Volume, Variety, Velocity, Veracity and Value.

**B.** 1. a. **Numerical Data:** Numerical data have meaning as a measurement such as a person's height, weight, IQ or blood pressure; or they're a count such as the number of stock shares

a person owns or how many pages you can read of your favourite book before you fall asleep. Statisticians also call numerical data as quantitative data.

b. **Categorical Data:** Categorical variables represent types of data which may be divided into groups. Some variables are categorical by the length of their unique values. For instance, if a variable has only unique values [-2, 4, 56], you could treat this variable as categorical. The colour of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd and terrier) would be examples of categorical variables.

c. **Ordinal Data:** Ordinal data mixes numerical and categorical data. Ordinal data are often treated as categorical, where the groups are ordered when graphs and charts are made. However, unlike categorical data, the numbers do have mathematical meaning. For example, if you survey 100 people and ask them to rate a restaurant on a scale from 0 to 5, taking the average of the 100 responses will have meaning. This would not be the case with categorical data.

2. Data collection methods can be classified as:

- **Primary data collection:** Since primary data is the data that is collected for the first time through personal experiences or evidence, especially for research, data collection methods for primary data are observations, surveys, personal interviews, telephonic interviews, case studies, etc.
- **Secondary data collection:** Secondary data is a second-hand data that is already collected and recorded by some researchers for their purpose. Data can be collected for secondary data through government publications, censuses, books, websites and reports, etc.

  Secondary data collection method is affordable, easily available, and saves cost and time. But one disadvantage is that the data collected is for some other purpose and may not meet the present research purpose or may be inaccurate.

3. Many different methodologies can be used for data collection and analysis. Some of these are:

- Interviews
- Surveys
- Focus groups discussions
- Medical records
- Questionnaires
- Case studies
- Phone records
- Statistical method

**Survey Method:** Survey is one of the common methods of diagnosing and solving social problems. Many research problems require systematic collection of data from population through the use of personal interviews or other data gathering devices.

4. Following are some of the advantages of focus group discussion:

- Clarify and test preconceived notions and findings.
- Helps seek the common views.
- Hear customers' feedback in their own words and voices.

- Uncover ideas and issues that initially may nay not have been considered but are important to the customer.
- Discover the decision making process.
- Have the flexibility to dive deeper into issues that come up during the discussion.

5. 
- Big data analytics allow us to monitor and predict the developments of epidemics and disease outbreaks. Integrating data from medical records with social media analytics enables us to monitor flu or corona outbreaks in real-time.
- Advertising and marketing agencies are tracking social media to understand responsiveness to campaigns, promotions and other advertising mediums.
- Social media can provide real-time insights into how the market is responding to products and campaigns. With those insights, companies can adjust their pricing, promotion and campaign placement on the fly for optimal results.
- Big data allows cities to optimise traffic flows based on real time traffic information as well as social media and weather data.

6. The following techniques may be adopted for interpretation and analysis of consumer data:
   - Use of a multiclass classification algorithm
   - Anomaly detection algorithms
   - Regression Algorithm
   - Clustering
   - Reinforcement learning
7. There are many different techniques for multivariate analysis and they can be divided into two categories:
   - **Dependence techniques:** Dependence methods are used when one or some of the variables are dependent on others. Dependence looks at cause and effect; in other words, can the values of two or more independent variables be used to explain, describe or predict the value of another dependent variable? To give a simple example, the dependent variable of 'weight' might be predicted by independent variables such as 'height' and 'age'.
   - **Interdependence techniques:** Interdependence methods are used to understand the structural makeup and underlying patterns within a dataset. In this case, no variables are dependent on others, so you're not looking for causal relationships. Rather, interdependence methods seek to give meaning to a set of variables or to group them together in meaningful ways.
8. The aim of multivariate analysis is to find patterns and correlations between several variables simultaneously. Multivariate analysis is especially useful for analysing complex datasets, allowing you to gain a deeper understanding of your data and how it relates to real-world scenarios.

## Higher Order Thinking Skills

Do it yourself

## Applied Project

Do it yourself

# 3. Data Visualizations

## Unsolved Exercise

### Objective Type Questions

**A.** 1. a 2. a 3. d 4. b 5. a

**B.** 1. T 2. T 3. F 4. F 5. T

### Standard Questions

**A.** 1. Data visualization techniques enable business analysts to understand which areas need to be improved. These techniques also enable to identify which factors control customer satisfaction and customer dissatisfaction. These techniques give a more detailed prediction and possible development for customers, company owners and decision-makers.

2. Trend analysis gives an idea to the traders based on what has happened in the past and what will be happening in the future. Trend analysis helps in predicting the future movement based on the current trending data.

3. For relatively small data sets where values fall into a number of discrete bins (categories), it is advisable to go for dot plots.

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

5. Number of occurrences of a particular data value in a data set is known as its frequency. For example, if four students have a score of 80 in mathematics, then the score of 80 is said to have a frequency of 4.

   The difference of maximum and minimum values is known as the range. The range is determined by subtracting the minimum value from the maximum value in a set of values.

**B.** 1. Data visualization is a technique to represent data graphically. Data visualization with five examples of real-life use of data visualization are:

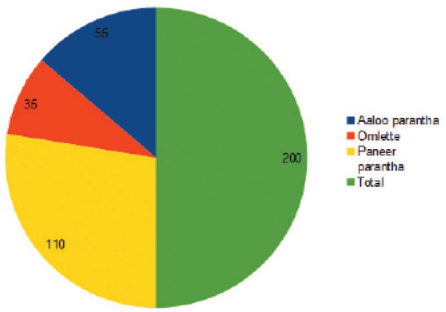
- Monitoring Progress of Students
- Identifying Trends in Business
- Identifying Usage Trend of a Website

- Monitoring Goals and Results of a Sales Executive
- Visualizing Spread and Impact of Pandemics

2. Pie chart, Bar chart, Histogram, Column Graph, Line Chart, Area chart, etc. are the names of few graphs/charts used for data visualization.
3. **Example of Single-variable plot:** A resident hostel of a particular college has 200 students. The hostel serves morning breakfast as per student's choice. On Sunday, hostel provides special menu in which they serve four items. The table below gives the details of student's choice for this feast:

| Breakfast Item | No. of Hostelers opted |
|---|---|
| Aaloo parantha | 55 |
| Omlette | 35 |
| Paneer parantha | 110 |
| Total | 200 |

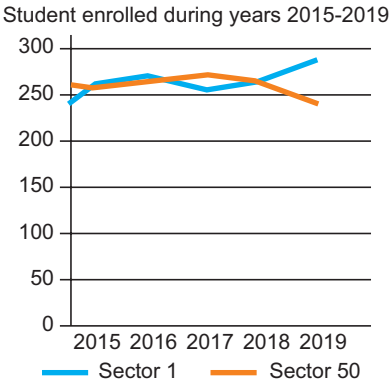
You can visulise the response by plotting pie chart very quickly:




**Example of multi-variable plot:** The following table shows the result of a survey of how many students enrolled in five years, i.e from 2015 to 2019 in two schools of Noida sector 20 and 50:

| Year of odd Semester | Number of students in sector 20 school | Number of students in sector 50 school |
|---|---|---|
| 2015 | 260 | 256 |
| 2016 | 268 | 266 |
| 2017 | 254 | 270 |
| 2018 | 264 | 260 |
| 2019 | 284 | 240 |

Following plot is created on the basis of the above data to visualise the data using a two-variable line chart:




4. a. Frequency distribution table

| Player | Runs | Frequency |
|---|---|---|
| P11 | 5 | 1 |
| P10 | 10 | 1 |
| P9 | 12 | 1 |
| P4 | 24 | 1 |
| P2, P7 | 32 | 2 |
| P6 | 35 | 1 |
| P5 | 43 | 1 |
| P1, P3, P8 | 45 | 3 |

b. Minimum value for runs scored is 5.

Maximum value for runs scored is 45.

## Higher Order Thinking Skills

Do it yourself

## Applied Project

Do it yourself

# 4. Ethics in Data Science

## Unsolved Exercise

### Objective Type Questions

**A.** 1. c 2. a 3. a 4. b

**B.** 1. T 2. T 3. F 4. T

### Standard Questions

**A.** 1. Data validity specifies that the values are according to the requirement. For example, ZIP codes are valid if they contain the correct characters for the region. In a calendar, months are valid if they match the standard global names.

2. Timeliness refers to the availability of data for analysis and for making decisions. It can be calculated as the time difference between when data is expected and when it is available for use.

3. Data governance helps to ensure that data is usable, accessible and protected. It also ensures that critical data is available at the right time to the right person, in a standardized and reliable form. Effective data governance leads to better data analytics, which in turn leads to better decision making and improved operations support.

4. One of the reasons many data governance initiatives fail is because the people involved have no clarity on their roles and responsibilities. This makes changes slower to implement and leads to many quality concerns being left unaddressed.

**B.** 1. Following are the key goals of ethical guidelines:

- Professional integrity and accountability: Professional integrity means to behave in accordance with ethical principles and act in good faith, intellectual honesty and fairness. Accountability is being responsible or answerable for an action.
- Data integrity: This can be defined as the reliability and trustworthiness of data. It specifies the state of your data that is valid or invalid or the process of ensuring and preserving the validity and accuracy of data.
- Informed consent: It is a procedure through which a customer after understanding all the information, can voluntarily provide his or her willingness to give personal details.
- Data confidentiality: It refers to protecting data against unauthorized access, disclosure or theft. It has to do with the privacy of data, including authorizations to view, share and use it.

2. The main goals and objectives of data governance include the following:

- To define, approve and communicate data strategies, policies, standards, architecture, procedures, and metrics.

- To track and enforce conformance to data policies, standards, architecture, and procedures.
- To sponsor, track and oversee the delivery of data management projects and services.
- To manage and resolve data related issues.
- To understand and promote the value of data assets.
- To improve internal and external communication.
- To increase the value of data for customers.
- To reduce costs of operation.
- To implement compliance requirements of agreements.
- To minimize business risks
- To establish internal rules for data use

3. Following are the benefits of improved data quality:
   i. **Accuracy of data:** Data accuracy is the first and crucial rule of the data quality framework. The accuracy of data provides a certain level of confidence to all who depend on that data.
   ii. **Completeness:** It is the second dimension of data quality that ensures there is no data missing from your data set and all possible data that is required is present.
   iii. **Uniqueness:** It is the most important dimension of data quality framework which specifies that there is no duplication of data.
   iv. **Validity:** It specifies that the values are according to the requirement. For example, ZIP codes are valid if they contain the correct characters for the region.
   v. **Consistency:** The data is said to be consistent if there are no conflicts in data within or between the systems which means that each user sees a consistent view of the data.
   vi. **Timeliness:** It refers to the availability of data for analysis and for making decisions.
4. Following are the primary barriers to data governance success:
   - **Weak data management framework:** A weak data management framework loses sight of how data is being used and shared.
   - **Prohibitions to data regulations:** It is becoming increasingly challenging to govern data despite data regulatory framework.
   - **Complicated data risk:** Despite applying data privacy and security measures, there is no guarantee that these are working efficiently and as intended. These measures do not ensure protection of data.

## Higher Order Thinking Skills

Do it yourself

## Applied Project

Do it yourself