Class VIII

ANSWER KEY

Data Science

1. INTRODUCTION TO DATA

Unsolved Exercise

Objective Type Questions

A. 1. c 2. a 3. c 4. d 5. c

B. 1. False 2. False 3. False 4. False 5. False

Standard Questions

A. 1. Qualitative data

2. No my height is not discrete data.
3. Data refers to raw facts and figures, while information is processed data that is meaningful and useful.
4. A calendar is an example of qualitative data.
5. Qualitative data

B. 1. The term 'data' is used to refer to the digital information or computer information. However, the data is an individual piece of information that contains raw facts about several things around you. These facts are available in many different forms like numbers, words, image, etc. This data is stored in the computer and can be transmitted from one computer to another. In most of the cases, the common forms of data include:
   - Text
   - Image
   - Speech or Sound
   - Video
   - Graphics
   - Spreadsheets

2. The data that has a finite value and can be counted is a sort of discrete data. This type of data generally involves use of integers. The valuWes of this type of data cannot be divided into

parts. For example, if a school has 30 students which is a fixed value, you can't have 30.5 as number of students. We will read in further classes that for purpose of computer the discrete data may assume binary, octal, decimal and hexadecimal values. Some of the examples of the discrete data are:

- Number of students in a class.
- Types of workers in a company: engineer, electrician, mason, plumber, carpenter, painter, helper, etc.
- Number of languages spoken in a country: Hindi, English, Bengali, Punjabi, Gujarati, Marathi, etc.
- Number of musical instruments: Piano, flute, guitar, violin, drum, etc.
- Number of provinces in a country.

3. The data that is measured or processed by machines as a continuous data series is called continuous data. The continuous data varies over a time period or spatially and can take any value between two numbers. Some of the examples of the continuous data are:
   - Amount of time required to complete a project.
   - Height of children.
   - Amount of time it takes to sell shoes.
   - Amount of rain, in inches, that falls in a storm.
   - Square footage of a two-bedroom house.
   - Weight of a truck.
   - Speed of cars.
4. If you have an account on Facebook, you might have noticed that you receive friend suggestions. That happens because, Facebook asks for your permission to access your contact information. This data is analysed to find many things like, if the people on your contact list are using Facebook or not. It also checks if they are connected to you on Facebook or not. If they are not connected to you, then Facebook recommends them to you. Isn't it strange how the Facebook knows your friends? Well, this is achieved through data accumulation and analysis. When you search for a friend on Facebook, the website uses your entered query as data and search in its database and then shows a list of people with same name. This is how data analysis is used in the social media websites in real life.

   Let us take another example of a cab booking website. The cab booking websites use a large amount of data of their drivers, vehicles, locations, trip fares, routes, etc. All this data is stored on their servers. When you book a cab, the website analyses its data and suggests you the cab types, fare and availability of cab, etc.

2. Introduction to Data Science

Unsolved Exercise

Objective Type Questions

A. 1. a 2. b 3. a 4. b 5. c 6. d

B. 1. False 2. True 3. False 4. True

Standard Questions

A. 1. Supervised learning describes a class of problem that involves using a model to learn a mapping between input values and the target variable. In supervised learning, we have to create a model for learning how to compare the input values with target variable and then find the difference.

2. Data mining is a process in which the data mining engineer collects huge amount of information from raw data lakes. The data lakes are the repositories where you can store large amount of data just like lakes can store rainwater in large capacity.

3. A very few organizations may require statisticians. They are required for developing an understanding of consumer behaviour and buying trends.

B. 1. Most prominent job titles in the field of data science are:

- Data Scientist
- Data Analyst
- Data Mining Engineer
- Data Architect
- Business Intelligence Analyst

Data Architect

Like building architects, the data architect has to work closely with various stakeholders of business organizations or projects. The stakeholders are users, owners, designers which have the actual blueprint and various codes that are to be complied with.

Data Mining Engineer

The data mining engineer collects huge amount of information from raw data lakes. The data lakes are the repositories where you can store large amount of data just like lakes can store rainwater in large capacity. Data is available to any authorised stakeholder. The data collection happens over a period of several years of organization's existence.

2. The job of a data scientist is to collect relevant data from various relevant sources available universally. A data scientist's role combines computer science, statistics and mathematics. He/she is responsible to store and organise the unstructured data. He/she converts the organised data into business solutions and finally, hands over the findings to business managers to run business positively.

3. You can ask certain questions to realise the objective of people coming to you for joining the business. You may ask questions in such a way that they may have only two options:

   Yes or No.

   Q: Whether you are looking for a new service or product?

   Q: Are you looking for sustainable product or only cheaper product?

   This type of query will help you to identify the type of customer. This mode is called classification.

   If your online transaction appears to be out of fraudulent activity, the value of transactions can be evaluated and checked against historical data available. Some examples are:

   Q: How much is the daily or monthly expenditure done by the customer concerned?

   A: Maximum 20 thousand on a particular day or 60000 thousand in a month.

   Q: Which types of accounts are held by Mr. A?

   A: Salary account or current account

   This type of researching is known as regression in data science.

4. If your online transaction appears to be out of fraudulent activity, the value of transactions can be evaluated and checked against historical data available. Some examples are:

   Q: How much is the daily or monthly expenditure done by the customer concerned?

   A: Maximum 20 thousand on a particular day or 60000 thousand in a month.

   Q: Which types of accounts are held by Mr. A?

   A: Salary account or current account

5. Anyone who has aptitude for statistical thinking and is fully conversant with information technology; he/she should also be creative and have curious mind with multi-modal communication skills and vast capacity for analysing can adopt a career as data scientist.

Higher Order Thinking Skills (HOTS)

1. Reinforcement learning means learning from a class of problems where a training agent creates an environment for getting feedback from the users and past experience of the organization.

   Flying taxies or drones are aerial vehicles that are envisioned to revolutionize urban transportation by providing efficient and eco-friendly alternatives to traditional ground-based modes of transit. RL plays a crucial role in enabling these flying taxis or drones to navigate dynamically changing environments, avoid obstacles, and make real-time decisions to ensure safe and efficient travel.

   For instance, Reinforcement learning algorithms can be used to train drones to optimize their flight paths based on factors like traffic conditions, weather patterns, and airspace regulations. By continuously learning from their interactions with the environment and receiving feedback

on their actions, these autonomous vehicles can improve their performance over time, leading to smoother and more reliable operations.

Furthermore, Reinforcement learning enables flying taxis or drones to adapt to unforeseen circumstances and handle complex scenarios, such as emergency landings or route adjustments due to airspace congestion. This flexibility and adaptability are crucial for the successful deployment of autonomous aerial transportation systems in urban settings.

2. One of the primary ways data science benefits sports teams is through performance analysis. By collecting and analyzing data on player performance, tactics, and opponent strategies, coaches and analysts can identify strengths, weaknesses, and areas for improvement. This information allows teams to tailor their training regimens, develop effective game plans, and optimize player positioning to optimize performance and achieve better results.

   Moreover, data science enables teams to enhance injury prevention and management strategies. By monitoring biometric data, player workload, and injury history, sports scientists can identify injury risk factors and implement targeted interventions to reduce the likelihood of injuries occurring. Additionally, data-driven rehabilitation programs can help injured athletes recover faster and return to peak performance sooner.

3. Data Visualisation

Unsolved Exercise

Objective Type Questions

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

B. 1. True 2. True 3. True 4. False

Standard Questions

A. 1. Histograms provide efficient methods using bars of different heights. Histogram is a graphical representation of the distribution of values in numerical data, grouping each value into a 'bin' and displaying the bin counts across the range of values.

2. Regression literally means state of returning to original value. But here in statistics, the regression analysis amounts to a finding of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).

3. Frequency distribution tables are made to show how often a group of data points appear in a given data set of observations made by you.

B. 1. You have to take care of following qualities of required data:

- Quality of Data: While gathering from different sources, you have to pay attention on quality of data. The poor quality of data collected may result in distortion in collected data set. Getting good quality of data from right type of source is of prime importance. You should always collect data from the sources of information which keep on updating themselves according to current trends. The data sources must be impartial and honest. The content available should be in direction of targets and fair enough.
- Completeness of Data: You should collect the information from the sources which have a complete set of observations, values and trends. Many a times while dealing with a particular package, you may have data without mentioning the timespan and without mentioning the purpose and target reader or listener or audience.
- Format of Data: Having a format of data helps the data analyser in a great way. First of all, the details should be available and formatted in such a way that is amenable to digital data processing. There are a number of formats for multimedia processing. The pictures and audio clips should be royalty free.

2. Following techniques of analysing data are very popular:

- Regression Analysis: Regression literally means state of returning to original value. But here in statistics, the regression analysis amounts to a finding of the relation between the mean value of one variable (e.g. output) and corresponding values of other variables (e.g. time and cost).
- Cohort Analysis: A cohort is a group of users (forum) experiencing a common product, event or service within the same time period. Cohort analysis is a kind of behavioural consumer (or any user) based analytics that break the data in a data set into related groups before analysis. These groups or cohorts usually share common characteristics or experiences within a defined time-span.
- Predictive Analysis: The science of analytics includes the discovery, interpretation and communication of meaningful patterns in data. It is especially valuable in areas rich in recorded information. Analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance.

3. Some types of analytics are better performed on some platforms than the others:

- Predictive analytics employs predictive modelling using statistical and machine learning techniques. Descriptive analytics such as reporting/OLAP (Online Analytical Processing), dashboards/scorecards and data visualisation have been widely used for some time, and are the core applications of traditional BI (Business Intelligence).
- Descriptive analytics means to look backwards (like at car's rearview mirror) and reveal what has occurred. One trend, however, is to include the findings from predictive analytics, such as forecasts of future sales on dashboards/scorecards.
- Decision analytics supports human decisions with visual analytics that the user models to reflect reasoning.

- Descriptive analytics gains insight from historical data with reporting, scorecards, clustering, etc.
- Prescriptive analytics recommends decisions using optimisation, simulation, etc.

Higher Order Thinking Skills (HOTS)

1. There are a lot of data visualisation tools like indicators, charts or graphs (bar, column, pie, area, etc.), tables, histogram, maps, etc.
   - **Indicators:** This type of visualisation scheme has been used very recently. This scheme is useful only when you want to display one or two numeric values which may be:
     - a number
     - a gauge type indicator
     - a status indicator
     - a ticker

   If you need to display one or two numeric values such as a number, gauge or ticker, use the indicator's visualisation. You can add additional features like title and a color-coded indicator icon such as a green up arrow or a red down arrow to represent the value in the clearest way.
   - **Charts or Graphs:** Charts or graphs are one of the most commonly used visualisation tools. The words charts and graphs are used interchangeably in common practice.
   - **Tables:** For comparative data analysis, data table or a spreadsheet is very efficient method. Representing your data through tables is a very old technique. They provide very compact and precise method of visualising the data at a place. In a two dimensional table, you may have several columns and rows. Normally, the items being compared are placed in columns, whereas categorical objects are shown in the rows.
   - **Histograms:** Histograms provide efficient methods using bars of different heights. They provide a great way to show results of continuous data, such as:
     - Weight
     - Height
     - Time required

   For statistical purposes, histogram allows you to see the frequency distribution of a data set. Histogram is a graphical representation of the distribution of values in numerical data, grouping each value into a 'bin' and displaying the bin counts across the range of values.
2. Do it yourself.
3. Below are some business intelligence software solutions of 2021 that come with the top data visualizations in the market:
   1. SAP BusinessObjects
   2. QlikView

3. Tableau
4. Cognos
5. Oracle BI
6. Sisense
7. Microsoft Power BI

Power BI is a visualization and analytics tool developed by Microsoft. It allows you to connect to a wide variety of data sources, design customized dashboards and detailed reports. It supports both mobile and web. Some of the advantages of Power BI are:

- Cloud-based
- Gives a single view of the dashboard
- Affordable
- Since this is a Microsoft tool, it has a very strong brand integration with the other MS tools
- A lot of documentation about this tool is available
- Big and active community
- A wide variety of charting options for data visualization are available
- Consistent upgrades
- Extensive database connectivity

4. Data Science and Artificial Intelligence

Unsolved Exercise

Objective Type Questions

A. 1. d 2. d 3. b 4. c 5. c 6. b
7. c 8. d 9. d 10 b

B. 1. True 2. True 3. True 4. False

Standard Questions

A. 1. A podcast can be defined as "A digital audio file made available on the Internet for downloading to a computer or mobile device, typically available as a series, new instalments of which can be received by subscribers automatically". Podcasts offer a convenient and portable way to consume audio content on the go, while web browsing provides access to a diverse range of multimedia content through a browser interface.

2. API stands for Application Programming Interface. APIs are a set of functions and procedures that allow for the creation of applications that access data and features of other applications, services, or operating systems.
3. Yes, VR can be considered as the use case of AI.

B. 1. Speech recognition or speech-to-text is a capability which enables a program to process human speech into a written format. It is also known as automatic speech recognition or computer speech recognition. We now use voice recognition technology in our everyday lives. Dictation is free online speech recognition software that will help you write emails, documents and essays using your voice narration. With this software, there is no longer need to type a long text. Initially, data science helps collect diverse audio data. Data science also optimizes these models for better performance and facilitates continuous improvement by integrating new data.

2. Image analysis is the extraction of useful information only from digital images and has applications in many fields from astronomy to zoology, including biology, medicine and industrial inspection.

   Images can tell anything and everything: from business to scientific insights about customers, to alerts about a faulty process like leakages in buildings and machines, judging authenticity of a document, and to analyse the patterns that can shape to be of a big value for a business decision.

   Following are the applications of Image Analytics:

   - Image analytics speeds up airport traffic.
   - Many of modern airports are acquiring upgraded technology that enables the use of biometrics such as finger or iris scanning as an alternative security screening measure.
   - Analyzing social media images for missing persons, for example, facial recognition technology is being used in Australia to identify missing persons.
   - Using image analytics for real-time crime investment and vehicle damage assessment through cameras.
   - Augmenting the capabilities of physical challenged persons.

3. Artificial Intelligence (AI) is a branch of Computer Science dealing with the construction of computational artifacts to carry out tasks in the real world, beyond the hermetic universe of numbers, instructions and machine data that flourish inside computers. Nowadays, the real world exists not just in our physical surroundings, but also in resources of text, pictures and other media that we create for one another, store and access electronically.

   The real-world artifacts of Artificial Intelligence researchers sometimes tackle extravagant tasks, like flying spacecraft, beating the world chess champion or performing some key human-like interactions with the world (as a specially-built full-scale humanoid robot).

   Artificial intelligence generally is an attempt to build machines that think as well as study mental faculties through the use of computational models.

Higher Order Thinking Skills (HOTS)

1. Data analytics thus provides automatic tool to process large volumes of online structured text data into quantitative data to generate wishful insights. Combined with the visualization tools, this technique enables business organizations to discover the story behind to make better unbiased decisions.

   Text analytics is also a process of drawing meaning out of written communication. In a customer experience context, text analytics means examining text that was written by or about customers. You find patterns and topics of interest and then take practical action based on what you learn.

   The text analytic techniques can be divided in following free technical areas:

   - Information retrieval
   - Data mining
   - Natural language processing (NLP)

   The words 'text mining' and 'text analytics' are often used interchangeably. The term text mining is generally used to derive qualitative insights from unstructured text, while text analytics provides quantitative results.

2. Natural language is any of the languages naturally used by humans i.e., not an artificial or man-made language such as a programming language. 'Natural language processing' (NLP) is a convenient description for all attempts to use computers to process natural language.

   NLP includes:

   - Speech synthesis: although this may not at first sight appear very 'intelligent', the synthesis of natural-sounding speech is technically complex and almost certainly requires some 'understanding' of what is being spoken to ensure, for example, correct intonation.
   - Speech recognition: it basically reduces continuous sound waves to discrete words.
   - Natural Language Understanding (NLU): moving from isolated words (either written or determined via speech recognition) to 'meaning'. This may involve complete model systems or 'front-ends', driving other programs by NL commands.
   - Natural Language Generation (NLG): generating appropriate NL responses to unpredictable inputs.