Grade VIII

DATA
SCIENCE

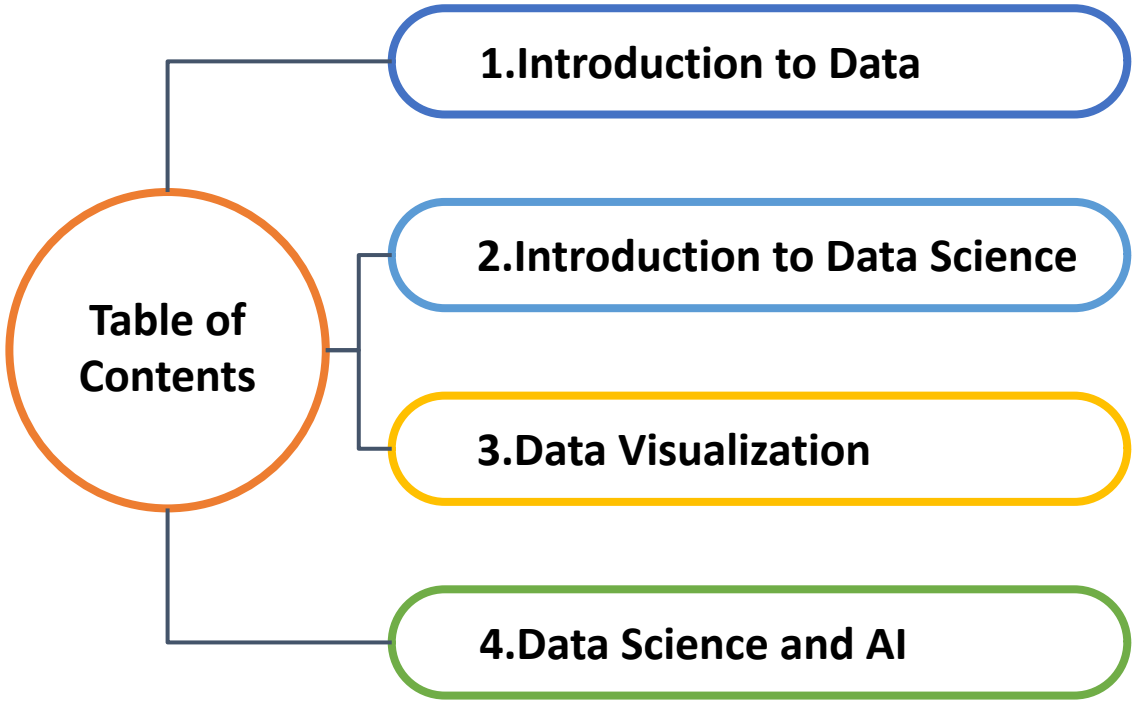


Table of Contents
1.Introduction to Data
2.Introduction to Data Science
3.Data Visualization
4.Data Science and AI

Chapter # 1

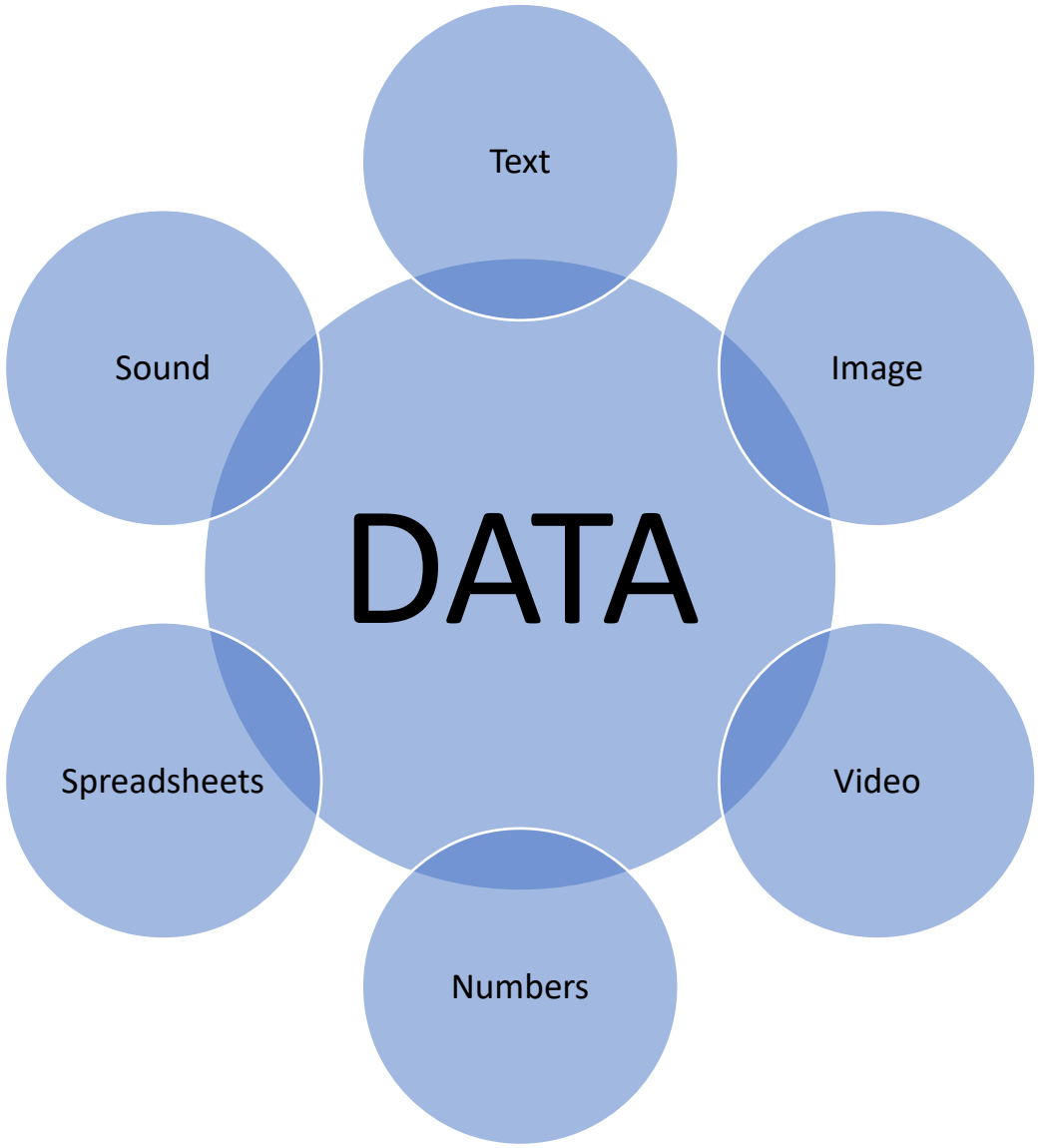
Introduction to Data

What is Data?

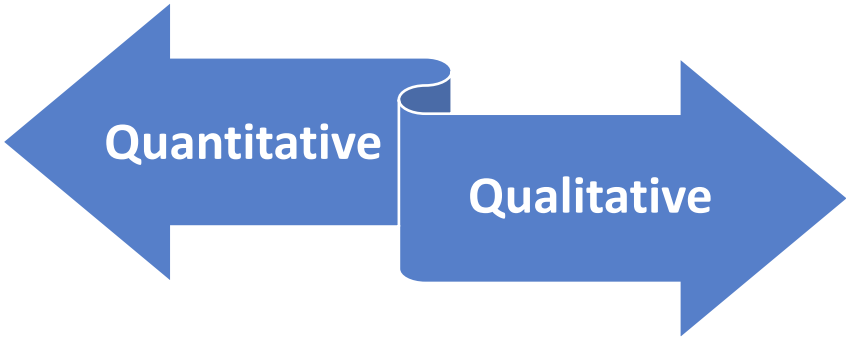
- ☆ We often use the term data to refer to **computer information.** This information is either **transmitted or stored**. Data comes in **numerous forms**.
- ☆ Any kind of information may it be in **numbers or text or pictures is termed as data**. Data is gathered and translated for some purpose, usually **analysis**.
- ☆ However, **if data is not put into context, it doesn't help** in any way to humans or computers.

Different forms of Data

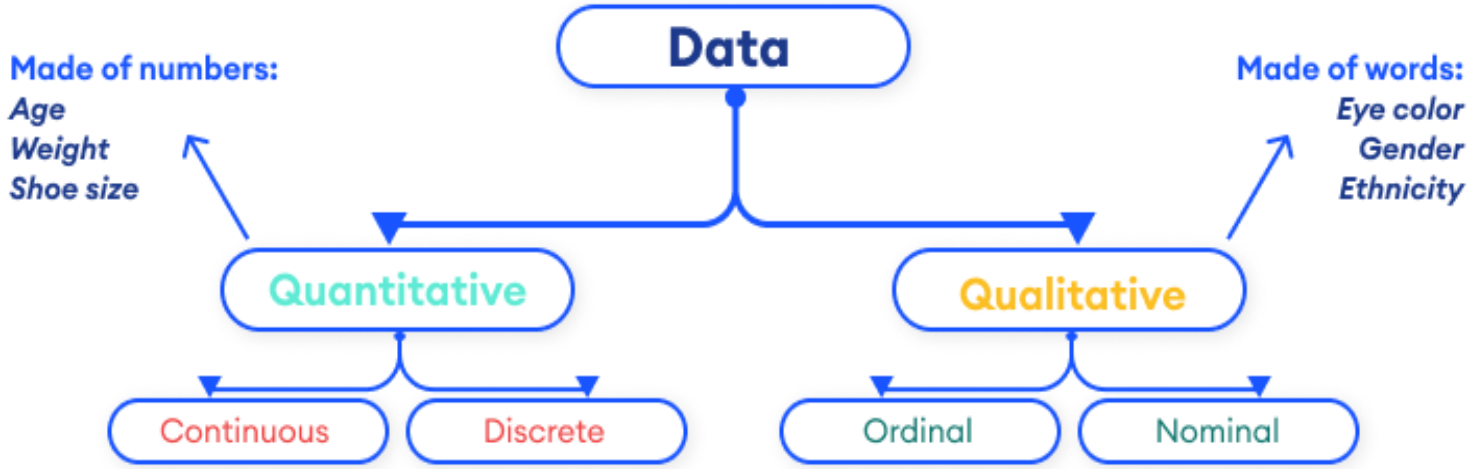
Data comes in different types. Internally in computers, data is stored as a series of **bits** that have a value of either **one or zero**.



Types of data in Statistics



1. **Qualitative** – Is a descriptive piece of Information.
For Example, “What a nice day it is”.
2. **Quantitative** – Data that is numerical information.
For example, “1”, “3.65” etc.,



Discrete Vs Continuous Data

EXAMPLE

Features of Discrete and Continuous data

Aspect	Discrete Data	Continuous Data
Definition	Consists of distinct values or categories	Can take any value within a given range
Nature	Countable, finite or countably infinite	Infinite and uncountable
Representation	Usually represented by whole numbers	Represented by real numbers
Examples	Number of students in a class, Number of cars in a lot	Height of students, Temperature, Weight of fruits
Measurement	Typically involves counting or enumeration	Requires measurement instruments (ruler, thermometer)
Graphical Representation	Bar charts, Histograms	Line graphs, Scatter plots, Density plots
Probability Distribution	Probability Mass Function (PMF)	Probability Density Function (PDF)

Real World - Example

	Continuous	Discrete		
	Quantitative data	Qualitative / Categorical / Attribute data		
Measurement	Units (example)	Ordinal (example)	Nominal (example)	Binary (example)
Time of day	Hours, minutes, seconds	1, 2, 3, etc.	N/A	a.m./p.m.
Date	Month, date, year	Jan., Feb., Mar., etc.	N/A	Before / After
Cycle time	Hours, minutes, seconds, month, date, year	10, 20, 30, etc.	N/A	Before / After
Speed	Miles per hour/centimeters per second	10, 20, 30, etc.	N/A	Fast / Slow
Brightness	Lumens	Light, medium, dark	N/A	On / Off
Temperature	Degrees C or F	10, 20, 30, etc.	N/A	Hot / Cold
<Count data>	Number of things	10, 20, 30, etc.	N/A	Large / Small
Test scores	Percent, number correct	F, D, C, B, A	N/A	Pass / Fail
Defects	N/A	Number of cracks	N/A	Good / Bad
Defects	N/A	N/A	Oversized, missing	Good / Bad
Color	N/A	N/A	Red, blue, green	N/A
Location	N/A	N/A	East, West, South	Domestic / International
Groups	N/A	N/A	HR, Legal, IT	Exempt / Non-exempt
Anything	Percent	10, 20, 30, etc.	N/A	Above / Below

2. Real world examples of Data

EXAMPLE

1 **Streaming Platforms (e.g., Netflix, Spotify):**

- ✓ ***Why they collect data:*** These platforms collect data on users' viewing/listening **habits**, **preferences**, and **interactions** with content to personalize recommendations and improve user experience.
- ✓ ***What they do with the data:*** They use algorithms to analyze user data and suggest **relevant movies**, **TV shows**, **music playlists**, or **podcasts**. This personalization enhances user engagement and retention.

2. Real world examples of Data

EXAMPLE

2 Social Media Platforms (e.g., Facebook, Instagram, TikTok):

- ✓ **Why they collect data:** Social media platforms gather data on users' ***demographics***, ***interests***, ***behaviors***, and ***interactions to tailor content***, ***ads***, and ***features***.
- ✓ **What they do with the data:** They utilize ***data analytics to target ads*** based on users' ***interests and behaviors***, ***optimize content visibility through algorithms***, and introduce new features based on user preferences.

2. Real world examples of Data

EXAMPLE

3 **Gaming Companies (e.g., Electronic Arts, Blizzard Entertainment):**

- ✓ ***Why they collect data:*** Gaming companies gather data on players' gaming behavior, preferences, and interactions to enhance gameplay, identify areas for improvement, and personalize gaming experiences.
- ✓ ***What they do with the data:*** They analyze player data to identify trends, optimize game mechanics, balance gameplay, and develop personalized content or in-game rewards tailored to individual players.

2. Real world examples of Data

EXAMPLE

4 Ticketing and Event Companies(e.g., Ticketmaster, Eventbrite):

- ✓ ***Why they collect data:*** Ticketing and event companies collect data on attendees' ***demographics***, ***purchasing behavior***, and ***event preferences to improve event planning***, ***marketing strategies***, and ***customer experiences***.
- ✓ ***What they do with the data:*** They use data analytics to ***forecast demand***, ***optimize ticket pricing***, ***target promotional campaigns***, and ***enhance event logistics*** based on attendee preferences and behavior patterns.

2. Real world examples of Data

EXAMPLE

5 **Movie Studios and Production Companies:**

- ✓ ***Why they collect data:*** Movie studios and production companies collect data on audience demographics, viewing habits, and feedback to inform marketing strategies, production decisions, and content creation.
- ✓ ***What they do with the data:*** They analyze audience data to identify target demographics, tailor marketing campaigns, optimize release schedules, and develop content that resonates with audience preferences, ultimately maximizing box office revenue and viewership.

Exercises

Objective Type Questions

1. Discrete data can take any value in a range (True/False).

False. Discrete data cannot take any value within a range; instead, it consists of *distinct values or categories*. Discrete data can only take on *specific*, *separate values*, often whole numbers, and cannot be subdivided into smaller parts. Examples ***include the number of students in a class, the number of cars in a parking lot, or the number of books on a shelf***.

2.Continuous data cannot take decimal values (True/False)

False. Continuous data can take *decimal values*. Continuous data represents *measurements* and can take any value within a given range, including fractions and decimals. Examples of continuous data ***include height, weight, temperature, and time. These values can be infinitely subdivided into smaller intervals, making them continuous.***

3. Information stored in a PDF is not considered data (True/False)

False. Information stored in a PDF is considered data. Data encompasses any collection of facts, figures, or other pieces of information that can be processed, stored, or transmitted. Whether it's text, images, tables, or other content, the information contained within a PDF document qualifies as data.

Exercises

Objective Type Questions

4. Quantitative data cannot take numerical values (True/False)

False. Quantitative data consists of numerical values that represent measurements or counts. This type of data is characterized by its ability to be quantified and expressed numerically. ***Examples of quantitative data include heights, weights, temperatures, and test scores. Therefore, quantitative data does indeed take numerical values.***

5. Qualitative data is descriptive in nature (True/False)

True. Qualitative data is descriptive in nature. It captures qualities, characteristics, or attributes that cannot be quantified numerically. This type of data is often obtained through observations, interviews, or open-ended survey questions and is used to understand attitudes, opinions, behaviors, and perceptions. ***Examples of qualitative data include descriptions of colors, textures, emotions, or opinions***.

6. “How is the weather like?” is what kind of data

Qualitative Data. It seeks information about the condition or characteristics of the weather, such as whether it's sunny, cloudy, rainy, windy, etc. This type of data provides descriptive information rather than numerical measurements or counts.

Exercises

Objective Type Questions

7. Which of the following is considered data?

a. Speech	b. Video	c. Messages	**d. All of the above**

8. How is the data used in the Entertainment Industry?

a. Predicting Interest	b. Targeting ads	**c. Both a & b**

9. Number of days in a week is an example of ?

a. Discrete Data	b. Continuous Data

10. What are the types of Quantitative data?

a. Discrete	b. Continuous	**c. Both a & b**

Exercises

Standard Questions

Please answer the questions below in no less than 100 words.

1. Explain what data is, with the help of two real-life examples.

Data refers to any ***collection of facts, statistics,*** *or* ***information that can be stored, processed,*** *or* ***analyzed***. It encompasses both quantitative and qualitative elements, providing insights into various phenomena. For instance, in retail - sales figures, customer demographics, and product preferences constitute valuable ***data used for market analysis and decision-making***. In healthcare- ***patient records, medical test results, and treatment outcomes serve as data for improving healthcare delivery and patient care***. Overall, ***data is the foundation for informed decision-making, problem-solving, and understanding patterns or trends in diverse fields***

2. How is the data Categorized?

Data is categorized based on its ***characteristics and nature***, typically into two main types: ***quantitative*** and ***qualitative***. ***Quantitative data consists of numerical measurements or counts***, ***allowing for mathematical analysis and statistical interpretation***. *Examples* ***include sales figures, test scores, and temperature readings***. On the other hand, **qualitative data comprises non-numerical information, describing qualities, attributes, or characteristics**. This type of data ***provides insights into attitudes, opinions, behaviors, and perceptions***. Examples ***include survey responses, interview transcripts, and observational notes***. Categorizing data into quantitative and qualitative types helps in organizing and understanding the information gathered from various sources for **analysis and decision-making purposes**.

Exercises

Standard Questions

Please answer the questions below in no less than 100 words.

3. What is Discrete Data?.

Discrete data consists of distinct, separate values that are countable and finite. ***These values are typically whole numbers or integers and cannot be subdivided further***. Discrete data often represents items that can be individually counted or enumerated, such as the number of students in a classroom, the number of cars in a parking lot, or the number of books on a shelf. It differs from continuous data, which can take on any value within a range. ***Discrete data is commonly analyzed using methods such as frequency distributions, histograms, and probability distributions, making it essential in various fields, including statistics, mathematics, and computer science***

4. What is Continuous Data?

Continuous data represents measurements and can take on any value within a given range. Unlike discrete data, which consists of distinct, separate values, continuous data can be infinitely subdivided into smaller intervals. Examples ***include height, weight, temperature, and time***. Continuous data is typically collected through instruments or devices that provide precise measurements, such as rulers, thermometers, or clocks. It is analyzed using methods such as probability density functions, scatter plots, and regression analysis. ***Continuous data is essential in fields such as science, engineering, economics, and environmental studies, where precise measurements and analysis of continuous variables are critical for decision-making and understanding complex phenomena***.

Exercises

Standard Questions

Please answer the questions below in no less than 100 words.

5. Give two examples of real-life applications of Data.

- ✓ **Real-time Traffic Management**: Data analytics is used to monitor ***traffic patterns, congestion levels, and accidents in urban areas***. By collecting data from traffic cameras, sensors, and GPS devices, transportation authorities can analyze traffic flow, identify bottlenecks, and adjust signal timings or redirect traffic to optimize road usage. This helps reduce travel times, minimize congestion, and improve overall road safety.

- ✓ **Personalized Medicine**: In healthcare, data analysis is revolutionizing treatment approaches through personalized medicine. ***By analyzing patient data, including genetic information, medical history, and lifestyle factors, healthcare providers can tailor treatment plans to individual patients' needs.*** This leads to more effective treatments, fewer adverse reactions, and improved patient outcomes.

Exercises

Higher Order Thinking Skills

Please answer the questions below in no less than 200 words.

1. How is data used by online streaming platforms.

- ✓ Online streaming platforms, such as Netflix, Spotify, and YouTube, *utilize data in various ways to enhance user experience*, *personalize content recommendations*, and *optimize business strategies*. Firstly, these platforms collect extensive data on user interactions, including viewing/listening habits, search history, ratings, and social media activity. This data is then analyzed using advanced algorithms and machine learning techniques to generate insights into user preferences, trends, and behavior patterns.
- ✓ One key application of data in streaming platforms is content recommendation. By analyzing user data, these platforms can suggest relevant movies, TV shows, music playlists, or videos tailored to individual tastes and interests. This personalization not only enhances user engagement but also increases user retention and subscription rates.
- ✓ Furthermore, streaming platforms use data to optimize content delivery and user interface design. They track metrics such as buffering time, video quality, and navigation patterns to ensure smooth playback and a seamless user experience across devices. Additionally, data analysis helps these platforms understand audience demographics, market trends, and content performance, enabling them to make informed decisions about content acquisition, production, and licensing.
- ✓ Overall, data plays a crucial role in driving innovation and competitiveness in the online streaming industry. By leveraging data effectively, streaming platforms can deliver a more personalized, intuitive, and satisfying entertainment experience to their users while also maximizing their own business success.

Exercises

Higher Order Thinking Skills

Please answer the questions below in no less than 200 words.

2. Give five examples of discrete data around you.

- ✓ **Number of Books on a Shelf**: The number of books on a shelf represents discrete data as it consists of distinct, countable values. You can easily count the number of books present, such as 10 books, 15 books, or 20 books, making it discrete.
- ✓ **Number of Students in a Classroom**: The number of students in a classroom is discrete data since it involves counting individual students. For example, a classroom may have 25 students, 30 students, or 35 students, each representing a distinct countable value.
- ✓ **Number of Cars in a Parking Lot**: The number of cars in a parking lot is discrete data as it represents a countable quantity. You can count the cars present in the lot, such as 50 cars, 100 cars, or 150 cars, making it discrete.
- ✓ **Number of Coins in a Purse**: The number of coins in a purse is discrete data since you can count the individual coins. For instance, a purse may contain 5 coins, 10 coins, or 20 coins, each representing a distinct countable value.
- ✓ **Number of Emails in an Inbox**: The number of emails in an inbox is discrete data as it involves counting the individual emails. For example, an inbox may have 50 emails, 100 emails, or 200 emails, each representing a distinct countable value.

These examples illustrate how discrete data consists of distinct, separate values that can be counted or enumerated.

Applied Project

Data analytics is revolutionizing the airline industry by enabling predictive models to anticipate and mitigate flight delays. Here's how it's applied:

- ✓ **Weather Condition Analysis**: Data analytics tools process vast amounts of weather data from satellites, radar systems, and meteorological stations to predict weather patterns that could impact flights. Machine learning algorithms analyze historical weather data and flight performance to forecast the likelihood of extreme weather events causing delays.
- ✓ **Route Restriction and Air Traffic Management**: Data analytics helps airlines optimize flight routes by considering air traffic congestion, airspace restrictions, and airport capacity. Predictive modeling assesses real-time air traffic data to identify potential bottlenecks and reroute flights accordingly to minimize delays.
- ✓ **Mechanical Delays Prediction**: Airlines use predictive maintenance algorithms to analyze data from aircraft sensors, maintenance records, and historical flight data to predict mechanical failures before they occur. By identifying potential issues in advance, airlines can proactively schedule maintenance to avoid flight delays.
- ✓ **Runway Availability Analysis**: Data analytics tools analyze runway utilization patterns, airport capacity, and historical runway maintenance schedules to predict runway availability. By forecasting runway closures or congestion, airlines can adjust flight schedules and minimize delays caused by runway limitations.

Overall, data analytics empowers airlines to proactively identify and mitigate factors contributing to flight delays. By leveraging historical and real-time data, predictive models enable airlines to optimize operations, improve efficiency, and enhance the passenger experience by minimizing the impact of delays.

Chapter # 2

Introduction to Data Science

Learning Outcome

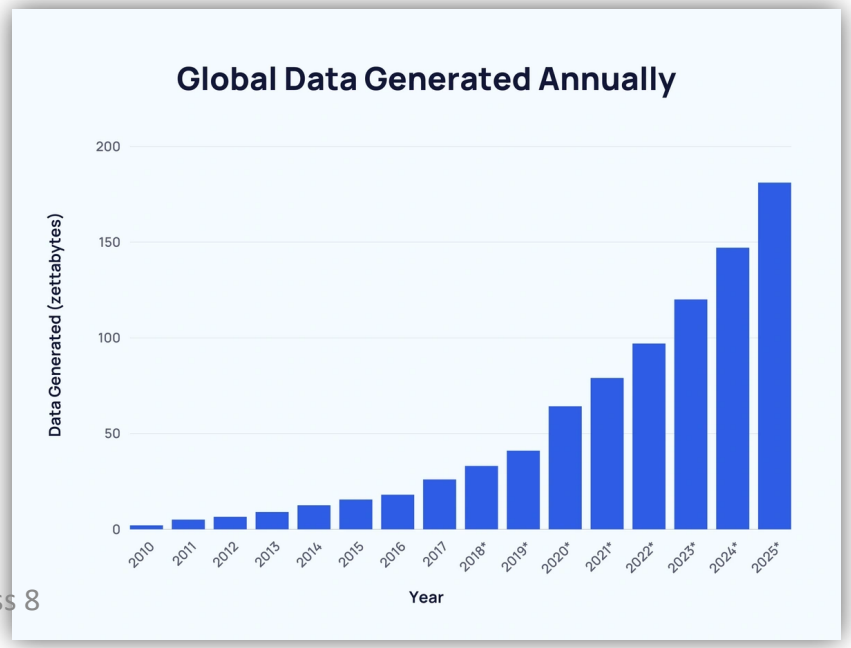
- ☆ What is Data Science?
- ☆ Careers in Data Science.
- ☆ What questions does Data Science answers?

1. A brief Introduction to Data Science

- ✓ Every day, ***our daily activities generate a vast amount of data***. Whether purchasing groceries, where each transaction is recorded, or withdrawing cash from various ATMs, each action is tracked to manage bank accounts accurately. Similarly, engaging with social media by liking posts or watching tutorial videos also contributes to data creation.

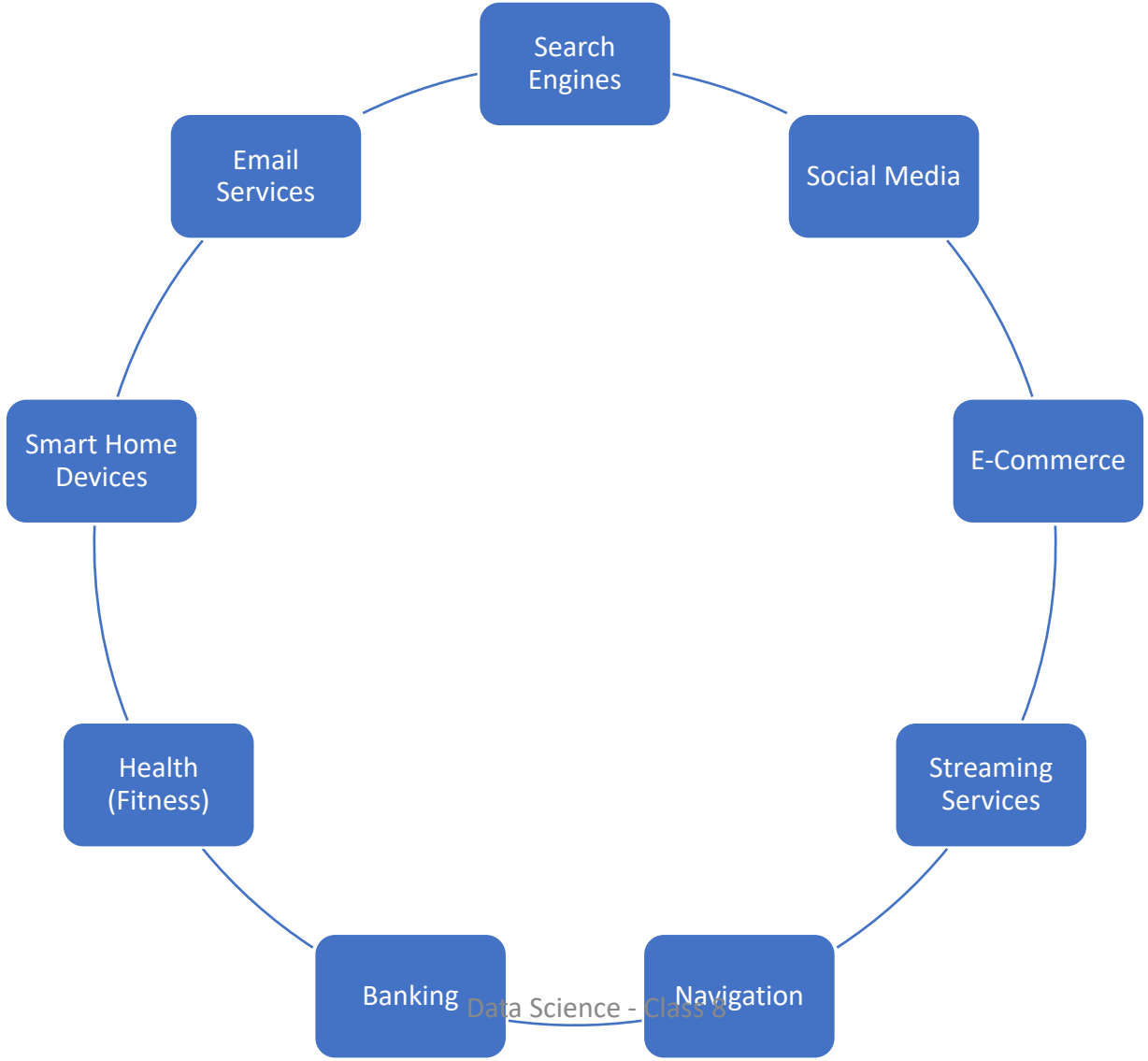
- ✓ This ***data can be analyzed to gain insights into our behaviors and preferences, potentially improving the offers and services provided to us***. This is the essence of data science: deriving meaningful insights from collected data to enhance decision-making processes. Data science applications are wide-ranging, from improving industry services to aiding law enforcement and enhancing sports training.

Unit of Measurement	Data Generated
Zettabytes	0.33
Exabytes	328.77
Petabytes	328,767.12
Terabytes	328.77 million
Gigabytes	328.77 billion
Megabytes	328.77 trillion
Kilobytes	328.77 quadrillion
Bytes	328.77 quintillion



Activity 2.1

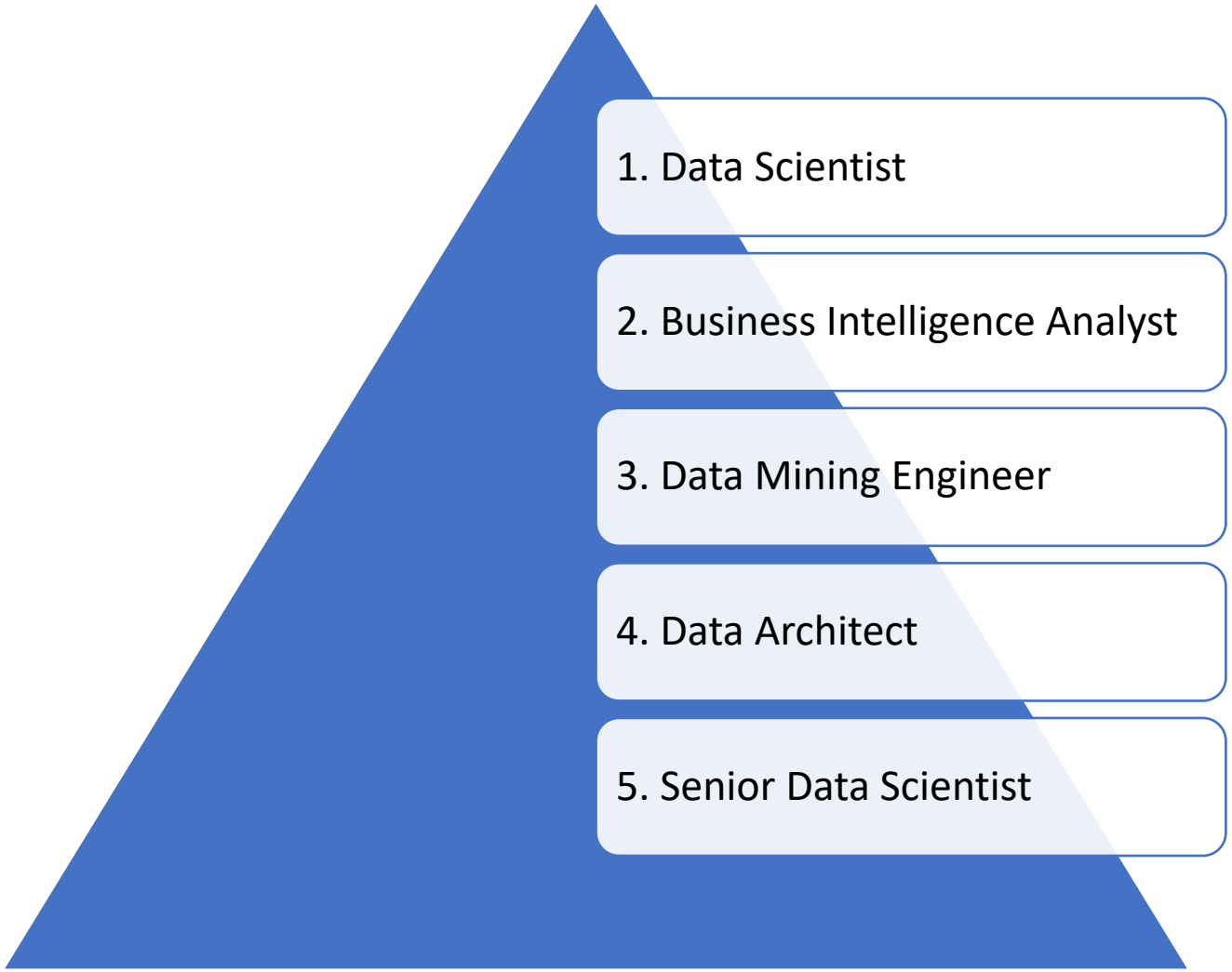
Try to find everyday used applications that depend on Data Science.



2. Careers in Data Science

The topic of what job options are available in data science becomes increasingly important as we dive deeper into the fields of data, data analysis, and data science. Our investigation has shown us the useful applications of data and data science across a range of industries. Some of us may have been inspired to seek a profession in this exciting field as a result of this.

Let's take a look at some of the Data Science career options to help you make an informed selection. Typical positions for someone in data science include:



1. Data Scientist - Data Scientists are passionate about data, specializing in the collection and analysis of both structured and unstructured data sets. This role blends elements of computer science, statistics, and mathematics. Data scientists process and model data, then analyze the outcomes to develop strategies that address company-specific needs. As analytical experts, they leverage their technological proficiency and insights into social science to identify patterns and oversee data management. Their deep understanding of industry-specific challenges allows them to devise tailored solutions to business problems.

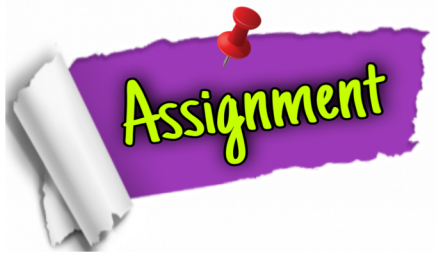
2.Business Intelligence Analyst - Business Intelligence Analysts use data to assess the market and find the latest business trends in the industry. This helps to develop a clearer picture of how a company should shape its strategy.

3.Data Engineer - Data Engineer examines not only the Data for their own business but also that of third parties. In addition to mining data, a data engineer creates robust algorithms to help analyze the data further.

4.Data Architect - Data Architects work closely with users, system designers, and developers to create a blueprint that data management systems use to centralize, integrate and maintain the data sources.

5.Senior Data Scientist - Senior Data Scientists anticipate the business's needs in the future. Although they might not be involved in gathering data, they play a high-level role in analyzing it. Using their vast experience, they can design and create new standards for analyzing data. They can also create ways to use statistical data and develop tools to further analyze the data.

Activity 2.2



Which career path would be good for you? Discuss

3. What does Data Science help us achieve?

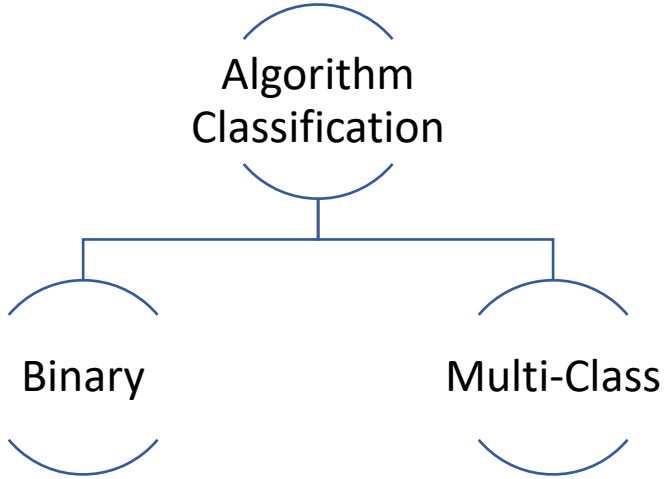
Data Science help us answer different types of questions that help us achieve various objectives.

Which class does this belong to –A or B?

The answers to some questions can only be from a definite number of options.

For example,
Q: ***Will it rain today?***
A: Yes/No

Q: ***Will the weather be hot or cold?***
A: Hot/Cold

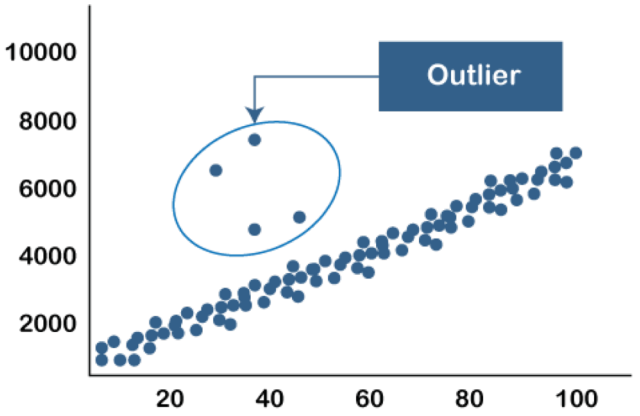


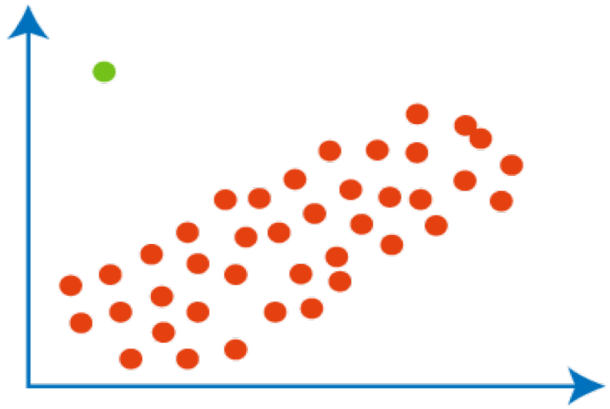
Binary classification involves categorizing data into one of two classes or categories. For example, determining whether an email is spam or not spam, predicting whether a patient has a particular disease or not, etc.

Multi-class classification involves categorizing data into more than two classes. For example, classifying images of fruits into categories like apple, banana, orange, etc.

What is ”Outlier”?

- The Objective is to find outliers or anomalies in data that is otherwise mostly consistent.
- These anomalies could be a cause of concern especially in cases where we need the data to be within a specific range all the time.
- An unexpected change in data patterns can often be a sign of something going wrong or possible fraud.
- Example: if an unexpected transaction is done from your debit card which not match your regular transactions, there could be a case of fraud. Banking institutions track these records and alert the customer that an unexpected transaction has happened, and this helps in protecting the customer’s money.





The algorithms that are used for these types of questions are called ***anomaly detection algorithms***.

Q: Is this email normal or spam?

Q: You are checking your car tyre pressure. Is the reading normal?

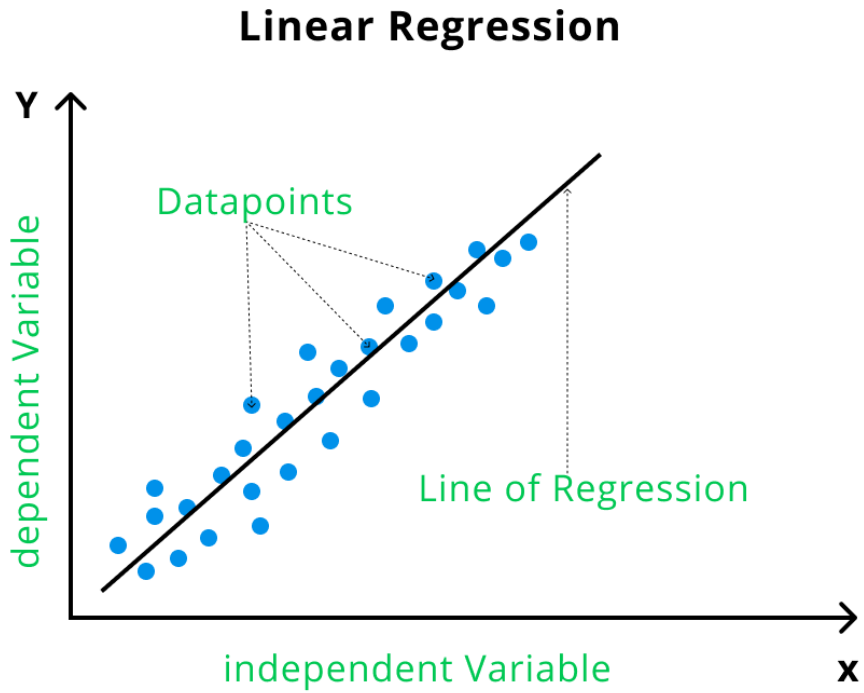
What is will probably be the value of this variable?

Machine Learning can also help us predict numerical values of continuous variables. There are scenarios in which we must predict numerical values of a variable based on historic data.

Some examples are:

Q: How much rainfall will we receive this year?
A: 100 mm

Q: How many runs will the winning team score?
A: 320



The kind of algorithms that can predict these values are called ***regression algorithms.***

How is the data grouped?

Data may be separated into distinct groups based on some parameters sometimes. This approach is called **Clustering** and is a type of **unsupervised machine learning**.

Examples:

Consider the data of the heights and weights of three species of cats. When we perform clustering, we will get an output that shows us the three different species of cats in three groups.

What should be done now?

This question usually solves the problems of autonomous robots or self-driving cars that need to make decisions based on changes in external factors. Machine learning helps to solve such problems with the help of ***reinforcement learning***.

These models are trained by a process of reward every time a correct action is taken and punishment every time a wrong action is taken. These systems are automated and can taken decisions without ***human intervention***.

How is the data grouped?

Data may be separated into distinct groups based on some parameters sometimes. This approach is called **Clustering** and is a type of **unsupervised machine learning**.

Examples:

Consider the data of the heights and weights of three species of cats. When we perform clustering, we will get an output that shows us the three different species of cats in three groups.

What should be done now?

This question usually solves the problems of autonomous robots or self-driving cars that need to make decisions based on changes in external factors. Machine learning helps to solve such problems with the help of ***reinforcement learning***.

These models are trained by a process of reward every time a correct action is taken and punishment every time a wrong action is taken. These systems are automated and can taken decisions without ***human intervention***.

Recap

- Data science is about how to extract meaningful interpretation from the data.
- There are many careers in Data Science like Data Scientist, Data Engineer and Data analyst.
- Data Architect and Senior Data Scientist are two roles for experienced professionals.
- Classification helps us to predict if a new item belongs to class A or class B.
- Regression helps us to predict the value of a continuous variable.
- Clustering helps us to find patterns in the data.
- Reinforcement learning helps models to take decisions based on external factors.

Exercise

Objective type Questions

Please choose the correct options in the questions below.

1. A school name ABC has recorded the total marks of every student in the class. This is an example of:

a. Qualitative *b. Quantitative* *c. Both* *d. None of the above*

2. A food delivery app has asked for your feedback on the quality of the food. You have written two paragraphs to describe the food. This is an example of:

a. Qualitative *b. Quantitative* *c. Both* *d. None of the above*

3. It would help if you predicted what the temperature would be for next Friday. Which algorithm will you use?

a. Clustering *b. Regression* *c. Anomaly detection* *d. Binary Classification*

4. You need to predict if you car tire will last for the next 1000 km. Which algorithm will you use?

a. Clustering *b. Regression* *c. Anomaly detection* *d. Binary Classification*

5. You want to build a way to segregate spam emails from good emails. Which algorithm will you use?

a. Clustering *b. Regression* *c. Anomaly detection* *d. Binary Classification*

Standard Questions

Please answer the questions below in no less than 100 words.

1. What are the common career paths for data science?

- *Common career paths in data science include **Data Analyst**, where you interpret and visualize data; **Data Scientist**, focusing on predictive modeling and machine learning; **Machine Learning Engineer**, specializing in algorithms and systems for AI; **Data Engineer**, who builds infrastructure for data generation, collection, and analysis; **Business Intelligence Analyst**, converting data insights into strategic business decisions.*

- *Additionally, roles like **Data Science Manager or Chief Data Officer** involve overseeing data-focused teams or departments, respectively, with pathways branching into specialized fields like NLP, AI research, or analytics consultancy. Each role requires a mix of technical skills and domain expertise.*

Standard Questions

Please answer the questions below in no less than 100 words.

2. What does a data architect do?

- *A Data Architect designs, creates, deploys, and manages an organization's data architecture.*
- *They define how the data is stored, consumed, integrated, and managed by different data entities and IT systems.*
- *Their responsibilities include developing database solutions, improving data quality, and designing strategies to resolve data issues.*
- *They ensure that the data systems are scalable, secure, and supportive of business goals.*
- *Data Architects work closely with IT teams, Data Engineers, and business units to create blueprints that align data management with technological advancements and business strategy, making data accessible and actionable for all stakeholders.*

Standard Questions

Please answer the questions below in no less than 100 words.

3. What are the difference between classification and regression?

- *Classification and regression are both types of **supervised machine learning**, but they differ primarily in their output. Classification predicts a discrete label, categorizing data into distinct classes (e.g., spam or not spam). It is used when the output is categorical.*
- *On the other hand, regression predicts a continuous quantity (e.g., house prices, temperatures). It is used when the output is a real or continuous value. Essentially, classification assigns data points to categories, while regression fits a model to predict numerical values based on input variables.*

Higher Order Thinking Skills (HOTS)

Please answer the questions below in no less than 200 words.

1. Discuss a recent innovation that makes use of reinforcement learning

- *Reinforcement learning (RL) is a type of machine learning where a computer program learns to make decisions by trying different actions and seeing what happens—sort of like learning to ride a bike by trial and error. A recent innovation using RL is in video games, where it helps non-player characters (NPCs) act more realistically.*
- *For example, NPCs can now learn from the player's actions. If a player often sneaks around rather than fighting, NPCs will start to patrol more carefully, making the game feel more challenging and dynamic. Unlike older games where NPCs had predictable behaviors, RL allows these game characters to adapt based on how you play, which makes the game different each time you play.*
- *This technology isn't just fun and games; it's also used in robotics, self-driving cars, and more. In each case, RL helps systems improve over time by learning from past experiences, much like how we learn from our own successes and mistakes. This makes machines smarter and more adaptable, capable of handling new and complex tasks in real-world environments.*

* **NPC** – *Non player character*

Higher Order Thinking Skills (HOTS)

Please answer the questions below in no less than 200 words.

2. Write a short note on how data science is helping sports teams.

- *Data science is like a super tool for sports teams, helping them play better and smarter. Imagine a coach who wants to find the best players for the team. Instead of just watching them play, the coach uses data science to analyze tons of information about each player's performances. This can include how fast they run, how accurately they shoot, or even how well they work with teammates.*
- *Using special software, data scientists can crunch these numbers to predict which players might score the most goals or prevent the most points against their team. They can also help coaches develop better training programs that are tailored to improve each player's unique strengths and weaknesses.*
- *Besides player performance, data science also helps in managing the health of athletes. By analyzing data from fitness trackers that monitor heart rate and other vital signs, teams can prevent injuries by catching signs of fatigue or stress early.*
- *So, data science isn't just about numbers. It's a way for sports teams to get a deeper understanding of their players and games, leading to smarter decisions on and off the field. This means better game strategies, improved player health, and hopefully, more wins!*

Applied Project

Emails are a part of daily communication. Sometimes we receive unwanted emails called spam. There are few techniques that email providers use to identify spam mails:

- Content-based filtering (**Analyzing the words**, **Occurrence**, **Distribution of words to identify spam mail**)
- Header filters (Reviewing the email header) Example: **Promo!**, **Offer!**

Provide 2 example each of words / phrases in email content & header which marks an email as spam. Explain in detail, how email providers make use of clustering to mark an email as spam. Also elaborate how email providers create an update the words / phrases to mark an email as spam.

Content-based filtering:

1. Words/Phrases in Email Content: "Congratulations!", "Win Now!"

- **Explanation**: Phrases like "Congratulations!" and "Win Now!" often indicate promotional content or lottery scams commonly found in spam emails. These phrases are analyzed along with other content features to detect patterns indicative of spam messages.

2. Words/Phrases in Email Header: "Special Offer!", "Act Fast!"

- **Explanation**: Phrases like "Special Offer!" and "Act Fast!" in the email header often signal marketing or promotional emails, which are frequently marked as spam by email providers due to their high volume and unsolicited nature.

Clustering for Spam Detection:

- Email providers use clustering algorithms to group similar emails together based on their content features such as words, phrases, and their distribution. For example, they might use k-means clustering to categorize emails into clusters based on similarities in content. By analyzing these clusters, the provider can identify patterns specific to spam emails, such as frequent occurrences of certain words or phrases.

Creation and Update of Spam Markers:

- Email providers continuously collect data on new spam emails reported by users and analyze them to identify emerging patterns or keywords. They may also collaborate with security firms to stay updated on the latest spam tactics. When a new pattern or keyword is identified, it is added to the provider's spam detection system, ensuring that future emails containing similar content are flagged as spam. This process involves a combination of automated algorithms and manual review to maintain accuracy and effectiveness in spam detection.

Chapter # 3

Data Visualization

Learning Outcome

- ☆ *What is data Visualization?*
- ☆ *The importance of visualization.*
- ☆ *Collecting relevant data?*
- ☆ *Asking the right question*
- ☆ *Predict an Answer*
- ☆ *Examples of data visualization*

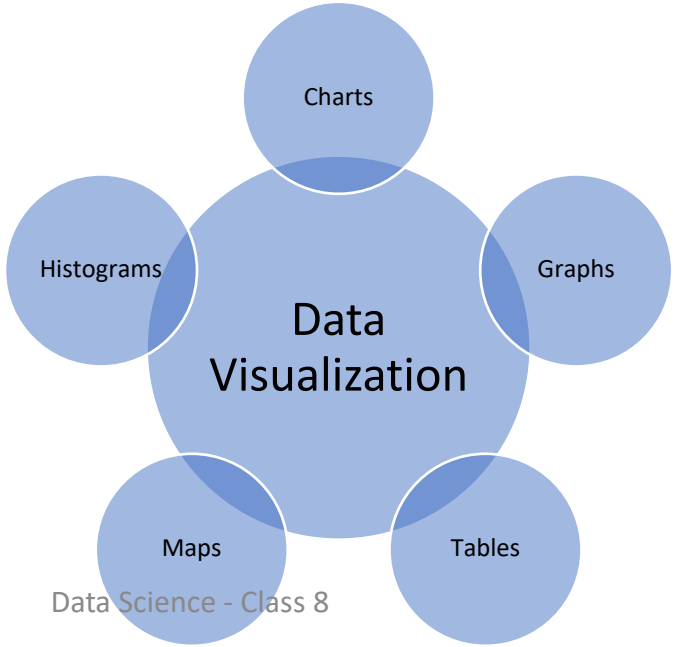
1. Introduction

- ✓ We have learned about how data is collect and how we can interpret the data by asking several types of questions on the data. Here, we will learn to visualize data and make predications.

2. What is data visualization?

- ✓ Data visualization is the representation of data or information in a graph, chart, or other visual formats.
- ✓ Data visualization provides a way to see and understand trends, outliers, and patterns in data
- ✓ Charts and graphs make communicating data findings easier even if you can identify the patterns without them.
- ✓ The goal of data visualization is to communicate information clearly and efficiently to users.

Data Visualization - Types

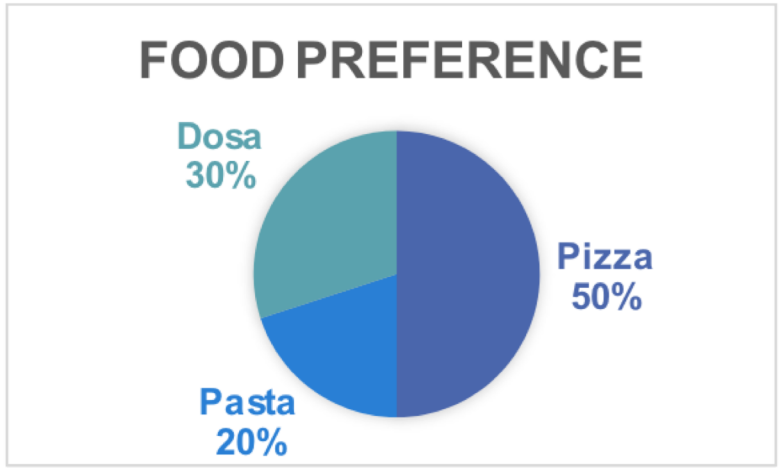


3. Examples of data visualization

- ✓ Example 1: Using a pie chart that displays the data of the food preferred by the students. We have the food item preference of 50 students. Let us now visualize the data using a pie chart and find the most preferred and the least preferred food item.

Food item	Number of students
Pizza	25
Pasta	10
Dosa	15

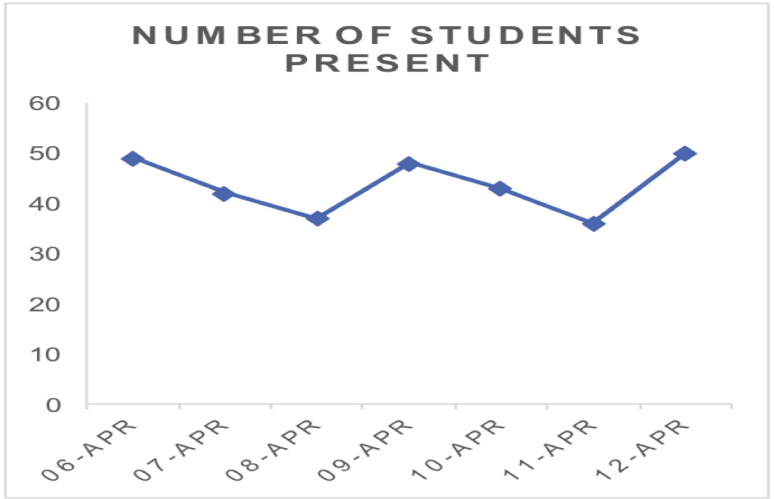
Visualize the data using a pie chart:



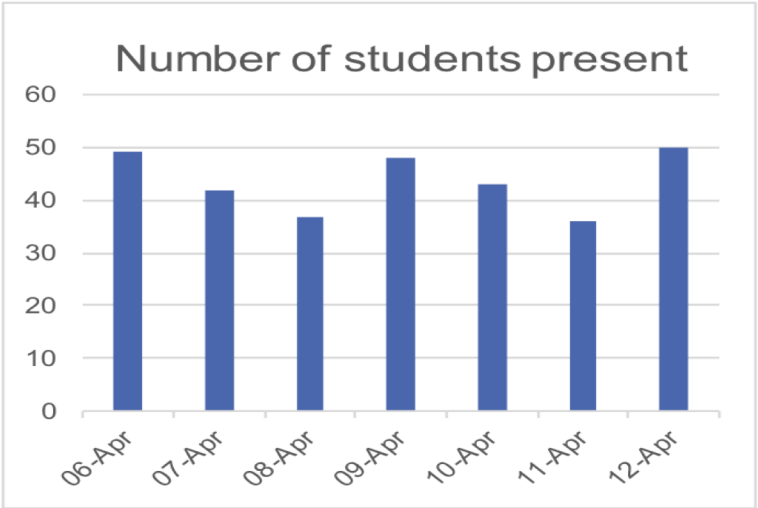
Example 2: Using a line chart that displays the data of the number of students present in the class for one week.

Here is the data:

Date	Number of students present
06-Apr	49
07-Apr	42
08-Apr	37
09-Apr	48
10-Apr	43
11-Apr	36
12-Apr	50



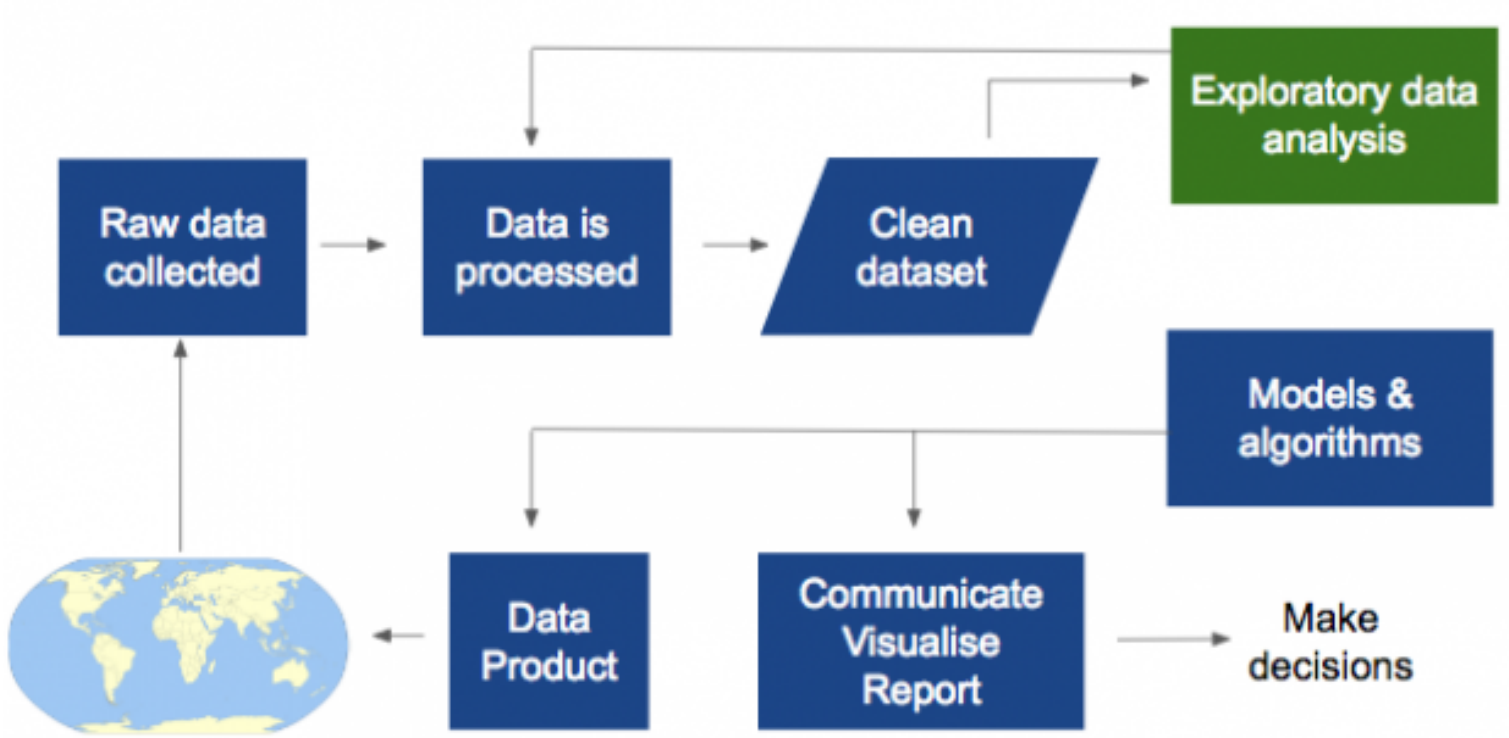
Visualize the data using a Line chart:



Visualize the data using a Bar chart:

4. Importance of data visualization

- To make sure that we get the ***required outcome from the data***, ***we must collect the right and relevant data***.
- It is essential to ***have correct and good quality data*** to make an analysis or to construct algorithms that can have an impact. Without ***relevant data, your analyses will not only be irrelevant, but they can also be misleading***.
- ***You cannot expect to find perfectly preprocessed raw data that be used directly for your needs***. Hence, you need to understand how the data was gathered and what sources it was collected from.



Let us understand what steps we need to take to make sure that we collect the right set of data for Analysis

- **Quality of the data** – Primary and most vital point to consider while collecting the data is the quality of data that is getting collected. If we collect incomplete data, build an unreliable database, and run analysis on skewed data sets, obviously we are not going to arrive at the required output. The quality of the data that is collected should always be the top priority while assessing the data.
- **Completeness of data** – We need to make sure that the data that is getting collected is a complete set. Incomplete sets of data may cause many discrepancies and wrong output on analysis.
- **Format of data** – The format of the Data that is collected for analysis should be right. Data should be accessible and readable for analysis. If the collected data is not in the right format, we should covert it to the required format for analysis.