Name - Piyusha R. Supe

Roll No - 2310315.

Assignment - 01

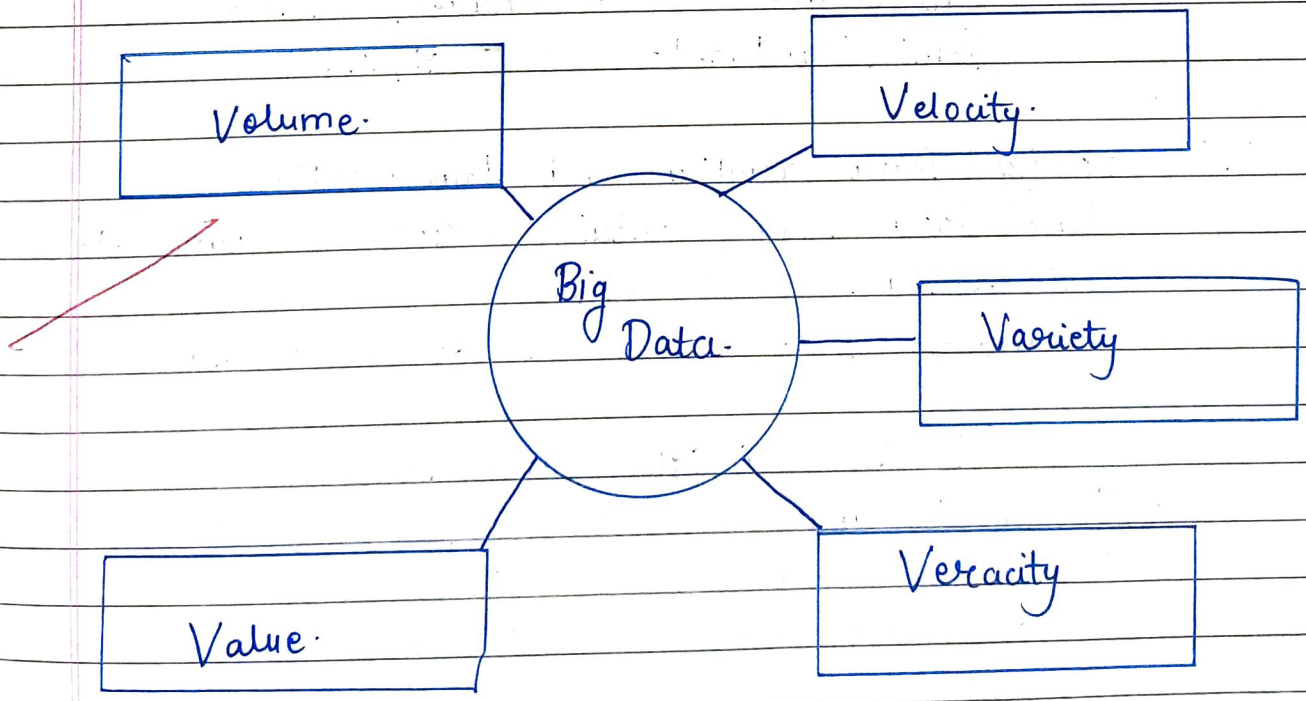
Based on Unit 1.

Q1. What is Big data? Explain characteristics of Big data?

→ 1. Big data refers to large complex datasets that cannot be processed using traditional data management tools due to their size speed and variety.

2. These datasets are generated using diverse sources including social media, IOT sensors, financial transactions, healthcare records and scientific research.

3. Characteristics of Big data -



4. Big data is typically described by -

a) Volume (size of data) -
Big data involves enormous amount of information, measured in terabytes (TB)., petabytes (PB) or even exabytes (EB).
Eg -
Facebook generates over 4 petabytes of data daily from user activities.

b) Velocity (Speed of data generation) -
Data is continuously generated at a high speed in real time or near real time.
Eg - Stock market transactions and live sports analytics generate data every second.

c) Variety (Different types of data) -
Data comes in multiple formats
- Structured (database tables)
- Unstructured (images, videos)
- Semi structured. (JSON, XML)

(d) Veracity (Data quality and trustworthiness.)
- Raw data may contain errors, duplicates, missing values, or biases.
- Example - Fake news on social media.

(e). Value (Business insights and Benefits) -
- The main goal is to extract useful insights that improve decision making.
- Ex - netflix analyses viewing patterns to recommend personalized content.

Q2. What is Big data analytics? Explain 5 Vs of Big data. Briefly discuss applications of Big data.

→ 1. Big data analytics refers to the process of analyzing large and complex dataset to discover hidden, patterns, trends, correlations, and insights. using techniques of -

- Machine learning.
- Artificial Intelligence.
- Data mining.
- Statistical modelling.
- Predictive analysis.

2. The 5 Vs of Big data are -

(a). Volume - size of data - Big data involves enormous amount of data measured in TB, PBI, EB.

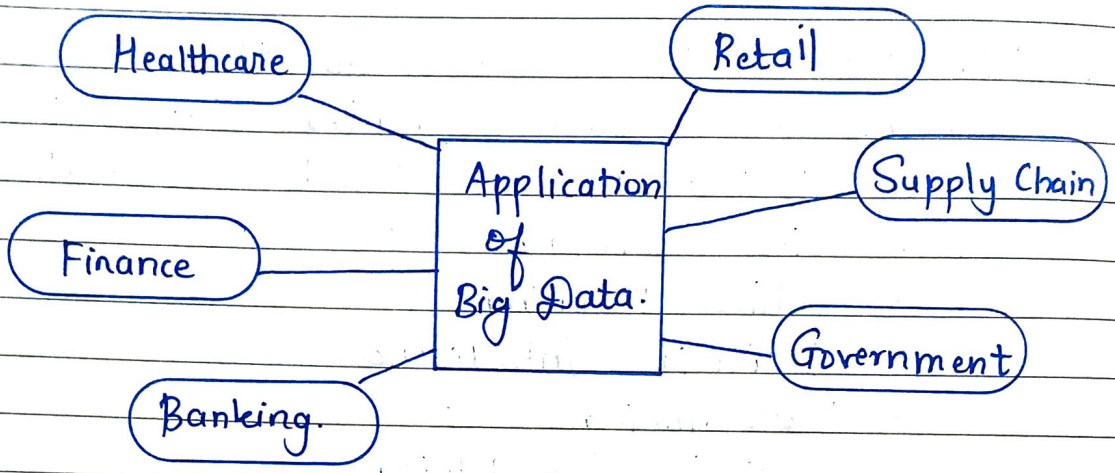
(b). Velocity - (Speed of data generation) - Data is continuously generated at high speed in real time.

(c). Variety. (Types of data) - Data comes in multiple formats.

(d). Veracity - Raw data may contain errors, duplicates and missing values.

(e). Value - The main goal is to extract useful insights that improve decision making. and analysis of data.

3. Applications of Big data –



1. Healthcare – • disease prediction using patient history.
   • Drug discovery using AI search.
   • Real time monitoring of ICU patients.

2. Finance and Banking – • Fraud detection using machine learning.
   • High frequency trading using real time stock market data.

3. Retail - ecommerce – • Personalized recommendations
   • Customer sentiment analysis using social media data.

4. Manufacturing and Supply chain –
   • Predictive maintenance in factories.
   • Optimized supply chain using IoT data

5. Government and smart cities –
   • Traffic optimization using IoT sensors
   • Crime prevention through predictive analysis.

Q3. What are Benefits of Big Data? Discuss challenges under big data. How big data analytics is useful for development of smart cities?

→ 1. Benefits of Big data –

a) Better decision making – Companies use Big data insights to make data driven decisions.

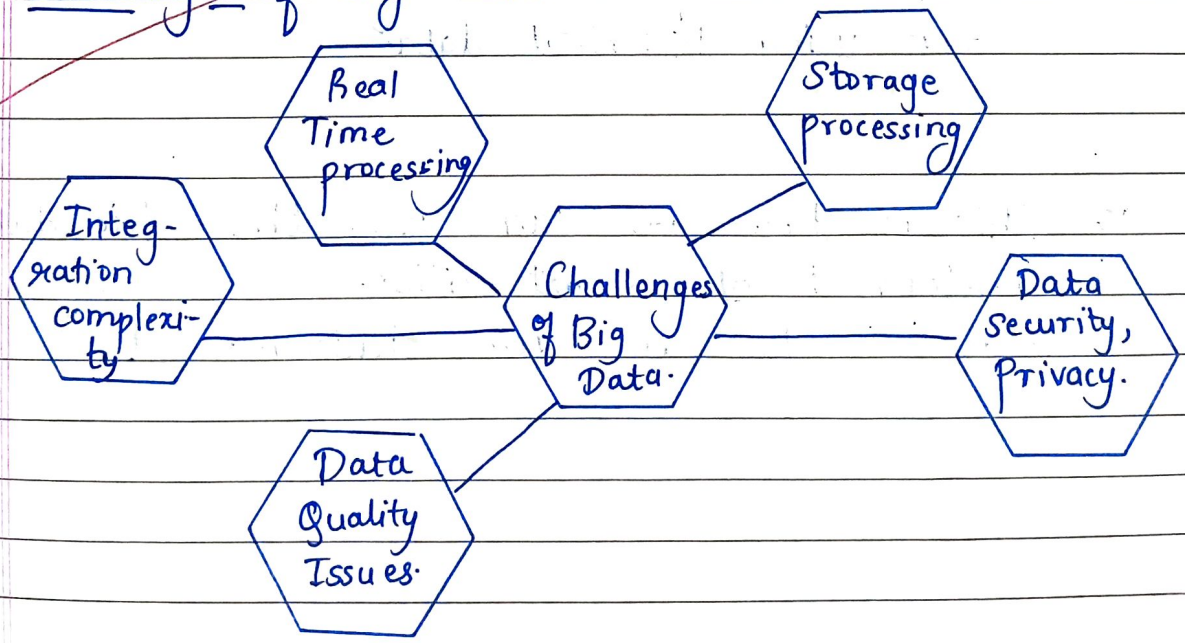
b) Frawd detection and cyber security – Detect anomalies in banking transactions.

c) Healthcare advancements – AI driven dignostics and precision medicine.

d) Improved Customer experience – Personalized recommendations (eg. Netflix, Amazon)

e) Cost Optimization – Efficient inventory and supply chain management.

2. Challenges of Big data –



3. Big Data for Smart Cities -

Big data plays a crucial role in developing smart cities by optimizing urban infrastructure.

a) Traffic Management -
- AI driven smart traffic signals.
- Google maps predicts traffic jams using real-time GPS data.

b) Waste Management -
- IoT enabled smart bins optimize waste collection.

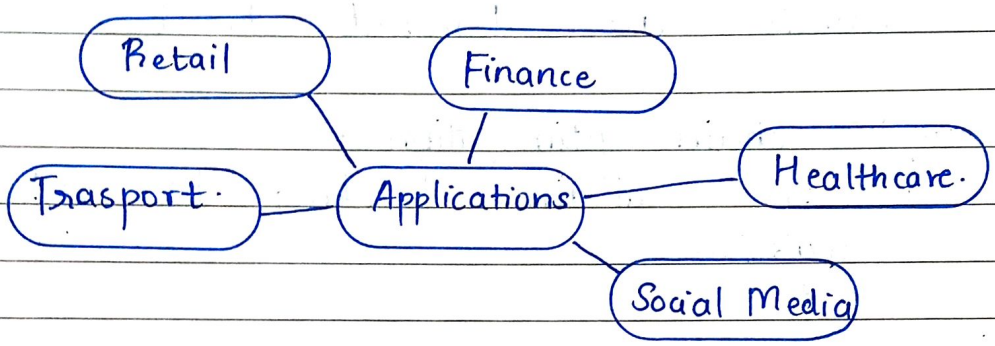
c) Energy Optimization -
- Smart grids adjust electricity distribution based on demand.

d) Crime Prevention -
- Predictive policing identifies crime prone areas using historical data.

- So this is how we can use the Big data in development of smart cities. Although it had a few challenges to resolve.

Q4. List various applications of Big data. How it can be used to improve business for Superstore -

→ I) Applications of big data -



1. Retail - Personalized shopping experiences, demand forecasting.
2. Finance - Risk analysis, fraud detection, algorithmic encoding.
3. Healthcare - AI driven diagnosis, hospital resource management.
4. Social Media - Sentiment analysis, targeted advertising
5. Transportation - Route optimization, fleet management

II) How Big data improves Superstore Business -

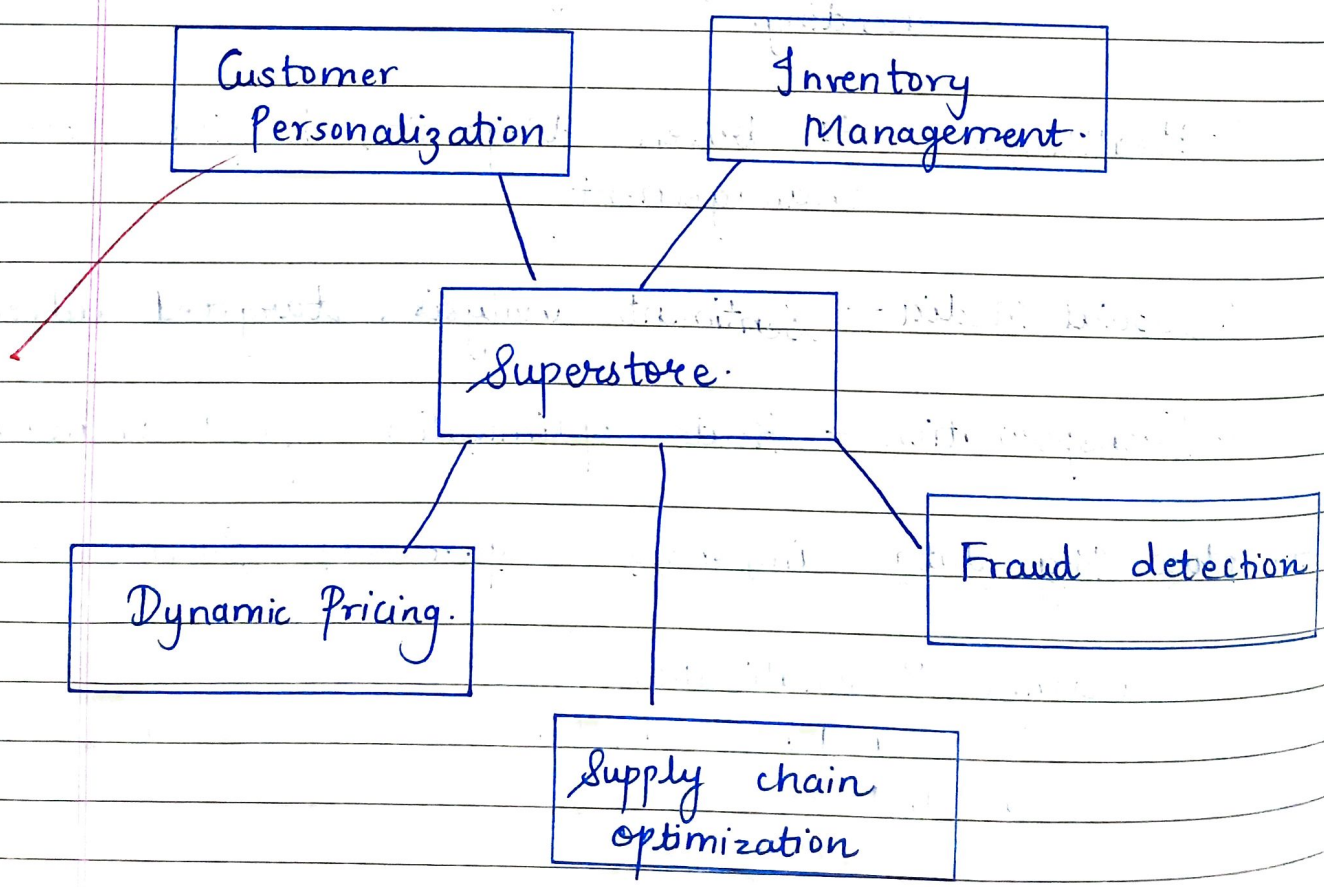
1. Customer Personalization -
   - Uses purchase history to recommend products.

2. Inventory Management -
   - AI predicts demand and prevents stock shortages.

3. Dynamic Pricing -
   - Adjusts prices based on competitor analysis and demand trends.

4. Supply chain optimization -
   - Uses real time tracking to optimise logistics.

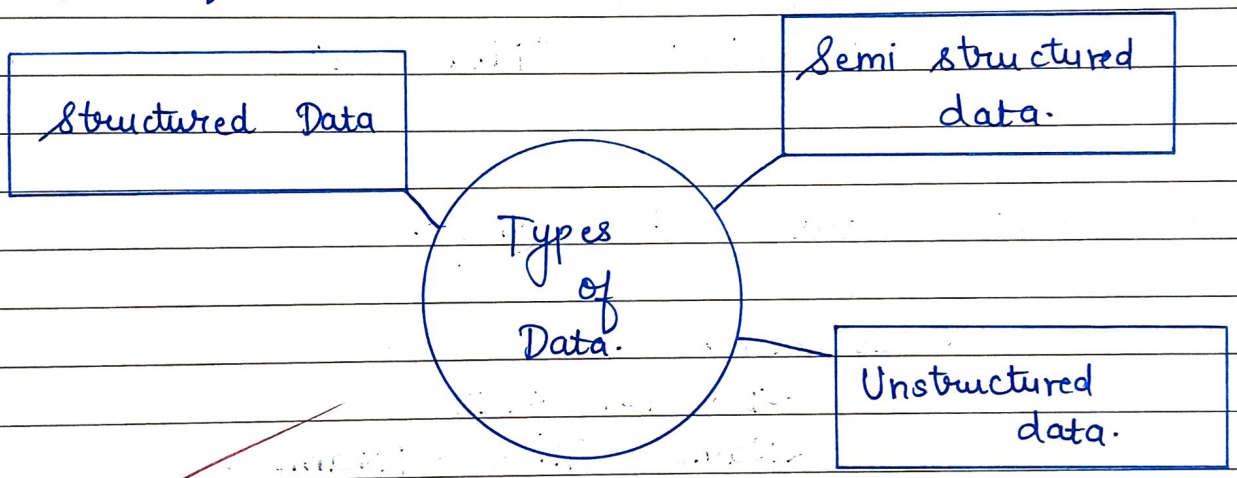
5. Fraud detection -
   - Detects unusual buying behaviour to prevent fraud.



Q5. What is data serialization? With proper example discuss and differentiate structured, unstructured, and semi structured data. Make a note on how many types of data serialization.

→ • Serialization is the process of converting data into a format that can be stored or transmitted and later reconstructed.
Eg - JSON, XML, Avro, Protocol Buffers.

• Types of data —



Data type	Definition	Example
1. Structured	Organized, tabular data with predefined schema.	SQL, Excel.
2. Semi-structured.	No strict schema but uses tags.	JSON, XML, log files.
3. Unstructured	No predefined format, difficult to analyse	Image, videos, emails, audio

- Examples of Serialization in different datatypes :-

1) Structured Data (SQL - JSON)
A relational database entry converted to JSON.

```
{
    "CustomerID" : 101,
    "Name"       : "Alice",
    "Age"        : 30,
    "location"   : "New York"
}
```

2) Semi structured data (XML).

```
<customer>
    <ID> 101 </ID>
    <Name> Alice </Name>
    <Age> 30 </Age>
    <location> New York </location>
</customer>
```

3) Unstructured - A jpg image converted into a binary data stream.

- Types of data Serialization -

1. Text based Serialization - Human readable (JSON, XML).

2. Binary Serialization - Efficient but not human-readable (AVRO, Protocol).

3. Hybrid Serialization - Combines text, and binary formats (Parquet, ORC).

Q6 How do you think of Big data analytics method as per your perspective of the following methods?

1. Data cleaning.
2. Data transformation.
3. Data integration.
4. Data discretization.
5. Data Reduction.

→ Big data analytics involves several preprocessing steps to ensure data is clean, consistent, and structured for meaningful analysis.

(1) Steps in Data cleaning -

- Handling missing data - Filling missing values with mean, median, mode or predictive models.
- Removing duplicates - Eliminating repeated records to avoid biased analysis.
- Correcting errors - Fixing incorrect data such as spelling mistakes, date errors or format inconsistency.
- Standardization - Ensuring consistent data format eg - 01-01-2024 to January 1, 2024.

(2) Data transformation.

- Normalization – Scaling values between 0 to 1 to standardize them
- Encoding Catagorical data – Converting catagorical to numerical data – (Male/Femele) – (0,1).
- Aggregating data – Summarizing hourly sales data into daily or weekly sales.

(3) Combining data from multiple sources into unified view – Data integration.

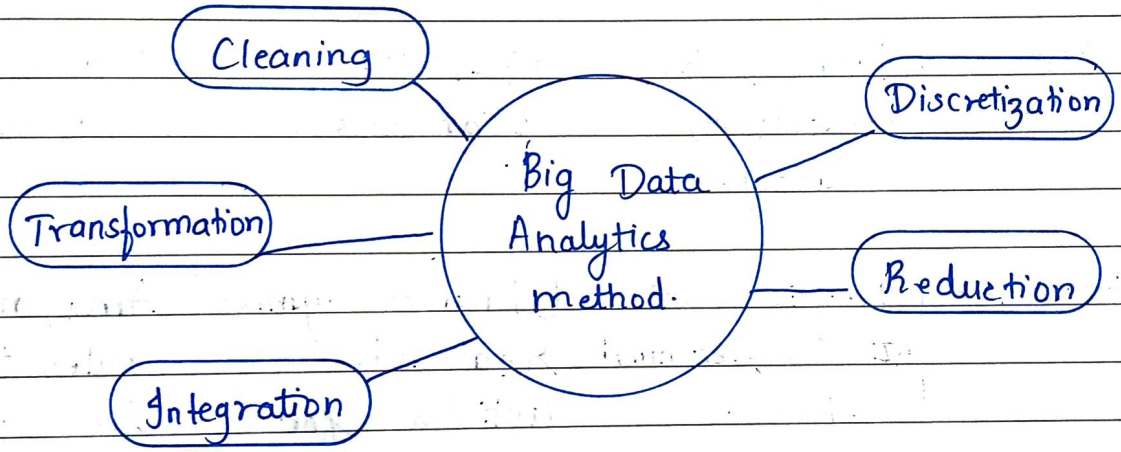
- Schema mismatch – Different databases have different formats
- Duplicate entries – Same customer name appears twice in meged dataset.
- Data inconsistency – Sales in USD vs INR currency match. mismatch.

(4) Data discretization –

- Equal width binning – Dividing data into equal intervals (eg. age groups 0-18, 19-35, 36-50)
- Equal frequency binning – ensuring each catagory has equal number of observation.
- Clustering – Using K means or heirarchical clustering to categorize data.

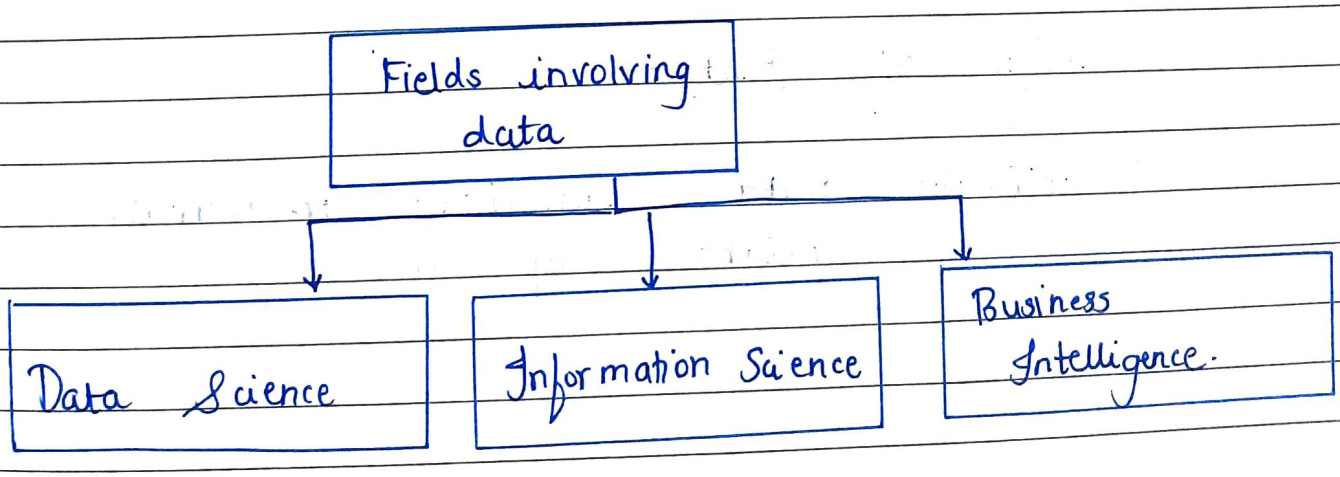
(5) Data Reduction –

- Dimensionality reduction : Using PCA (Principal Component analysis to remove irrelevant features.
- Sampling - Selecting a subset of data instead of processing entire dataset.
- Aggregation – Grouping data.



Q7 Define and compare data science, information science, and business intelligence. with example.

→



	Aspect	Data Science	Information Science	Business Intelligence
1.	Definition	Extracts insight using AI, ML, and analytics.	Studies how information is collected stored accessed.	Uses past data to support business decisions.
2.	Data Type.	Structured, Semi structured, unstructured.	Structured and Semi - structured	Primarily structured
3.	Tools Used.	Python, R, Tensorflow, Hadoop.	SQL, Data Warehouses.	Power BI, Tableau.
4.	Example -	Netflix uses AI to recommend shows.	Libraries organize research articles for easy access	Amazon analyse sales trends to optimize inventory.

Q8 Explain in detail about ~~d~~ nature of data and its applications -

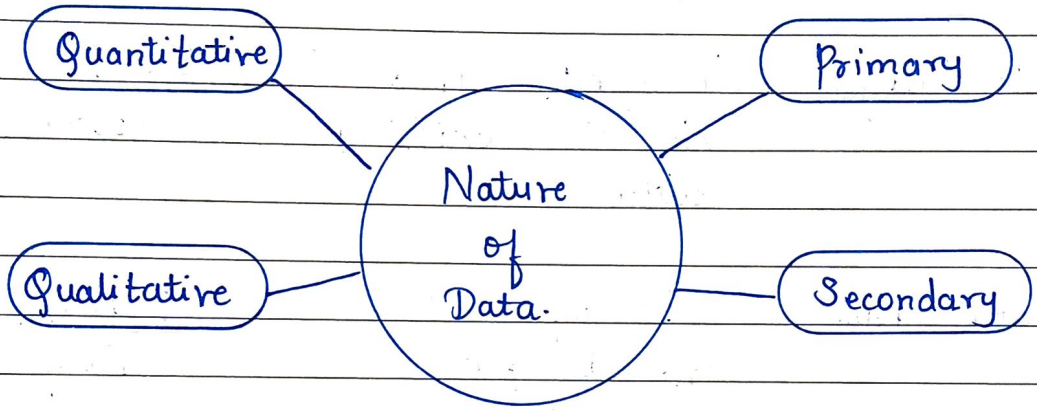
→ **Nature of Data -**

1. Quantitative Data - Numericals, measurable, (eg height, temperature)

2. Qualitative data - Descriptive catagorical (eg- gender, eye color)

3. Primary Data - Collected first hand (eg- surveys)

4. Secondary data - Pre-collected for another purpuse (eg - census data).

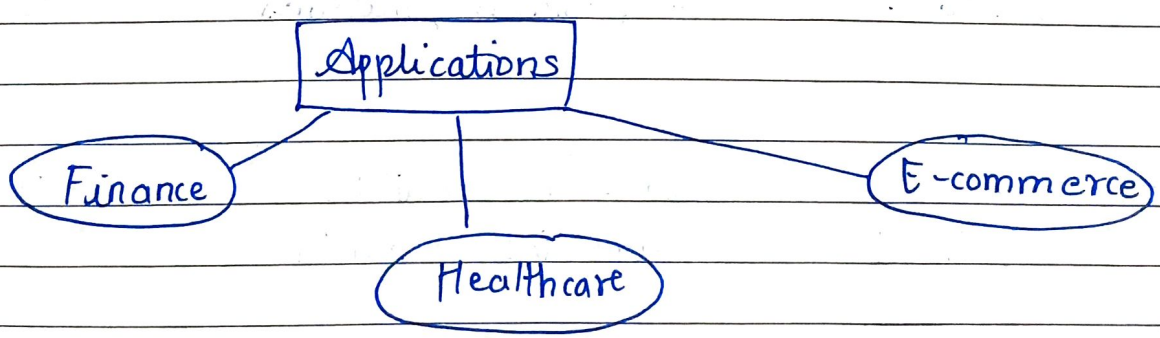


- Applications of Data -

1. Finance - Fraud detection, stock market prediction.

2. Healthcare - Disease prediction, electronic health records.

3. E-commerce - Recommendation systems, customer segmentation.



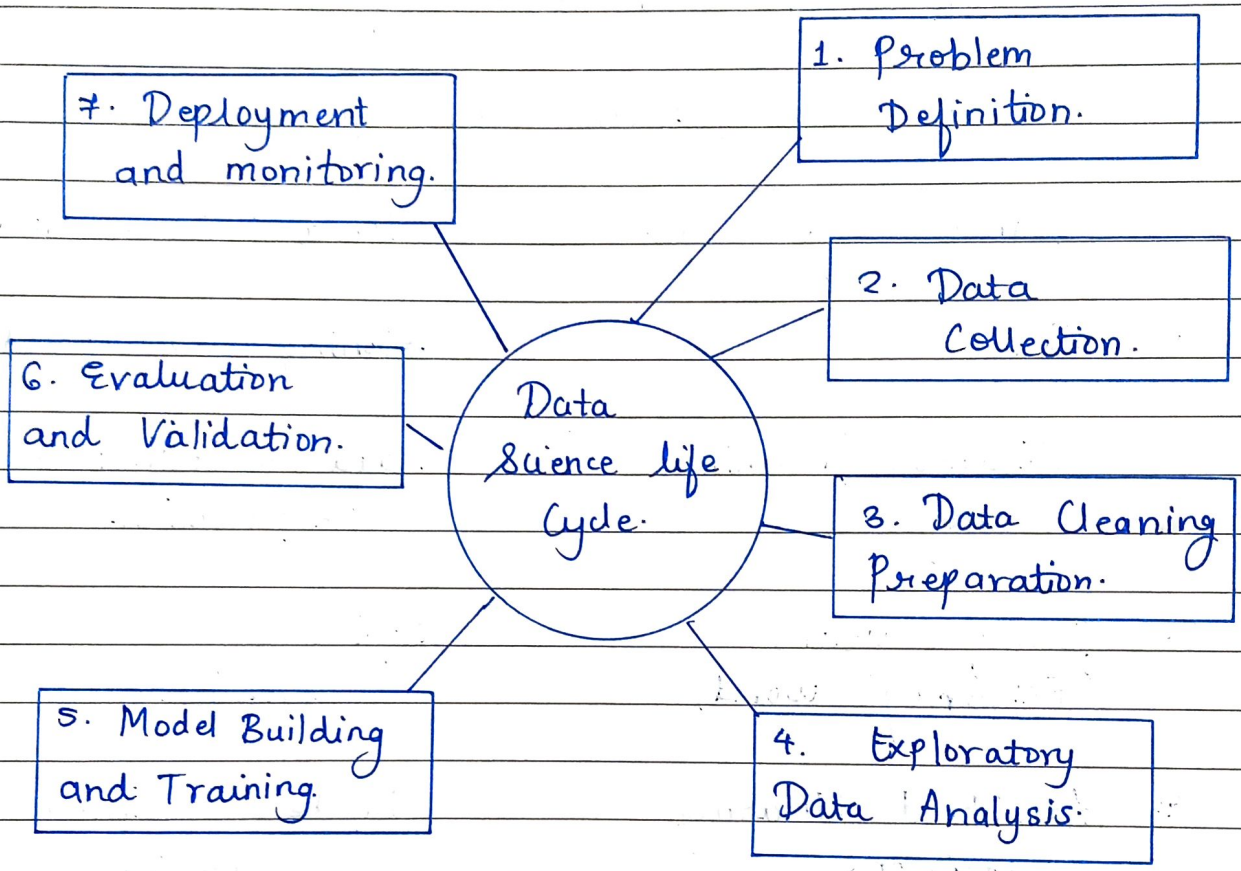
Q9. What are difference and compare with following terms: data analyst, scientist, data, administrator, statistician, business analyst, data architecture.

→

	Role	Responsibilities.	Skills Required.
1.	Data Analyst.	Interprets and visualizes existing data	Excel, SQL, Power BI.
2.	Data Scientist.	Develops ML models for predictions.	Python, AI, Deep learning.
3.	Data Administrator.	Manages database ensures security.	SQL, Database Management.
4.	Statistician.	Applies mathematical methods to analyse trends.	Probability, Hypothesis Testing.
5.	Business Analyst.	Translates data insights to business strategies.	Communication, Business Intelligence.
6.	Data Architect.	Designs and builds, data infrastructure	Hadoop, Data Modeling

Q10. What is data science life cycle? Relate it to human life cycle.

→ Data Science Life Cycle –



1. Problem Definition - Understanding the problem.
2. Data collection - Gathering relevant data.
3. Data cleaning and preparation - Removing noise and errors.
4. Exploratory Data Analysis - Identifying patterns.
5. Model Building and Training - Dereloping AI/ML models.

6. Evaluation and validation – Testing model performance.

7. Deployment and monitoring – Implementing the model in real world applications.

* Analogy to Human life Cycle –

	Human life Cycle:	Data Science life Cycle
1.	Birth.	Defining the problem.
2.	Childhood (learning stage).	Data collection and preparation.
3.	Teenage years (Exploring world)	Exploratory data Analysis.
4.	Adulthood (Making decisions)	Model Training and testing.
5.	Old age (Reflection and improvement)	Model deployment and monitoring.

Example –

- A baby is born (raw data)
- A child learns (data pre-processing and exploration)
- Adult makes decisions (model training and deployment)
- In old age adjustments are made (continuous improvement of AI models)

Name -
Piyusha R. Supe (23CO313).

RANKA
DATE / /
PAGE 19

Assignment - 02.

Based on unit 2.

Q1. Statistical hypothesis.

- H: $\mu = 75$ cents, where μ is true population average of daily per student candy + soda expenses in U.S. high schools.
- H: $p < 10$, where p is population proportion of defective helments for a given manufacturer.
- If μ_1 and μ_2 denote true average breaking strengths of 2 different types of twine, one hypothesis might be assertion that $\mu_1 - \mu_2 = 0$ another statement is $\mu_1 - \mu_2 > 5$.

→ A statistical hypotheses is an assumption or claim about a population parameter (such as the mean or proportion) that can be tested using statistical methods.

Lets analyze each hypotheses -

1. H : $\mu = 75$ cents.
   - Here μ represents true population mean of daily per student expenses on candy and soda in U.S. high schools.
   - This hypotheses claims that average spending is exactly 75 cents.

- It is a two tailed hypothesis (if tested) where deviations in either directions (higher or lower) could lead to rejecting H.

2. $H: p < 10$.

- Here, 'p' represents the true population proportion of defective helmets produced by a manufacturer.
- This hypothesis claims that less than 10% of helmets were defective.
- This a one tailed hypothesis (left tailed) because we are only interested whether defect rate is less than 10 percent.

3. $H: \mu_1 - \mu_2 = 0$ or $H: \mu_1 - \mu_2 > 5$.

- μ_1 and μ_2 represent true average breaking strengths of two different types of twine.
- The hypothesis $\mu_1 - \mu_2 = 0$ means there is no difference in breaking strengths between two types of twine.
- The hypothesis $\mu_1 - \mu_2 > 5$ means that twine type 1 is at least 5 units stronger than twine type 2.
- The first is null hypothesis (H_0) and second is an alternative hypothesis (H_a) right tailed test.

Q2 Null vs. Alternative Hypothesis -

Suppose a company is considering putting a new type of coating on bearings that it produces. The true average wear life with current coating is known to be 1000 hours. Which is denoting true average life for new coating the company would not want to make any (costly) changes until an evidence strongly suggested that it exceeds 1000.

→ In hypothesis testing, we always have:

- Null hypothesis (H_0): A statement of no difference or no effect.
- Alternative Hypothesis (H_a) : (or H_1) : The statement that we want to test and provide evidence for.

* Step 1 : Defining the hypothesis -

- Null hypothesis (H_0) : The new coating does not increase the average wear life.

$H_0 : \mu = 1000.$

- Alternative Hypothesis (H_A) : The new coating increases the average wear life.

$H_A : \mu > 1000.$

* Step 2: Explanation of the test.

- The company wants to avoid unnecessary costs by changing the coating only if there is strong evidence that new one is better
- This is a one tailed test (right tailed) since we are only interested in whether the new coating increases the lifespan.
- If statistical evidence supports Ha, the company may switch to new coating.

Q3. Test statistics -

Company A produces circuit boards but 10% of them are defective. Company B claims that they produce fewer defective circuit boards.

$H_0: p = 10$ versus $H_a: p < 10$.

Our data is random sample of $n = 200$ boards from company B. What test procedure or rule could we devise to decide if null hypothesis should be rejected?

→ • A test statistic is a value that helps determine whether to reject H_0.

- lets us understand the problem first -

- Problem –
  - Company A produces circuit boards and 10 % of them are defective.
  - Company B claims to produce fewer defective boards.
  - A random sample of 200 boards is taken from company B to test their claim.

* Step 1 : Define Hypothesis

- Null Hypothesis (H_0) : Company B has the same defect rate as company A.

$$H_0 : p = 0.10.$$

- Alternative Hypothesis (H_a) : Company B has a lower defect rate.

$$H_a : p < 0.10.$$

* Step 2 : Compute the test statistic –

- The proportion z-test is used –

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 (1 - p_0)}{n}}}$$

where,

$\hat{p}$ = sample proportion of defective boards from Company B.

p_0 = 0.10 (Company A's defect rate).

n = 200.

$$Z = \frac{\hat{p} - 0.10}{\sqrt{\frac{0.10\,(1 - 0.10)}{200.}}}$$

If $Z < -1.645$ (at 5% significance level), we reject H_0 and conclude Company B has a significantly lower defect rate.

Q4. Critical Region.

- The Brinell scale is measure of how hard it the material is. An engineer hypothesis that mean Brinell score of all sub critically noted ductile iron pieces is not equal to 170.
- The engineer measured brinell score of 25 peices of this type of iron and calculated sample mean to be 174.52 and sample standard deviation to be 10.31.

  Perform hypothesis test that true average Brinell score is not equal to 170 as well as corresponding confidence interval Set alpha = 0.01.

→ • Given data - :

- Sample mean $\bar{x} = 174.52$
- Hypothesized mean $\mu = 170.$
- Sample standard deviation 's' = 10.31
- Sample size $n = 25.$

- Significance level alpha = 0.01.

Step 1 : Define Hypothesis.

- Null Hypothesis (H_0) : $\mu = 170$.
- Alternative Hypothesis (H_A) : $\mu \neq 170$ (two tailed test).

Step 2: Compute test statistic (t-test).

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{174.52 - 170}{10.31/\sqrt{25}} \qquad t = \frac{4.52}{2.062} = 2.19.$$

Step 3 : Find critical t-value

- For $\alpha = 0.01$, $df = 24$, the critical t-values are ± 2.797
- Since, $2.19 < 2.797$, we fail to reject H_0.

Hence, this is how Brinell scale helps us determine scenario.

→

Q5 The first four raw moments of distribution are 2, 136, 320, 40000. Find co-efficient of skewness and Kurtosis.

→ • Given - first four raw moments.

$m_1 = 2$, $m_2 = 136$, $m_3 = 320$, $m_4 = 40000$.

• **Step 1** : Compute Variance (σ^2)

$$\mu_2 = m_2 - (m_1)^2 = 136 - 4 = \boxed{132}$$

• **Step 2**: Compute Skewness.

$$\mu_3 = m_3 - 3m_1m_2 + 2m_1^3 =$$

$$= 320 - 3(2)(136) + 2(8) = \boxed{-480}$$

$$\text{Skewness} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-480}{132^{3/2}} = \boxed{-0.1004}$$

• **Step 3** : Compute Kurtosis.

$$\mu_4 = 40000 - 4(2)(320) + 6(4)[136) - 3(16)$$

$$= 37656.$$

$$\text{Kurtosis} = \frac{\mu_4}{\mu_2^2} - 3 = \frac{37656}{17424} - 3 = \boxed{-0.84}$$

Hence,

Skewness = -0.1004 (Slightly left Skewed)

Kurtosis = -0.84 (platykurtic, meaning a flatter distribution).

Q6. Find the rank correlation co-efficient of the following data -

x	10	12	18	18	15	40
y	12	18	24	24	50	25

→ • Rank co-relation measures the degree of association between two ranked variables using Spearman's rank correlation co-efficient (r_s).

Step 1: Assign ranks.
We rank the values in ascending order for both -

x	Rank of x.	y	Rank of y.
10	1	12	1
12	2	18	2
15	3	24	3.5
18	4.5	24	3.5
18	4.5	25	5
40.	6	50.	6.

Step 2: Compute differences in rank -
$di = R(x) - R(y)$

X rank	Y rank	di	di^2
1	1	0	0
2	2	0	0
3	3.5	-0.5	0.25
4.5	3.5	1	1
4.5	5	-0.5	0.25
6	6	0	0

$$\Sigma d_i^2 = 1 \cdot 5$$

Step 3: Compute rank corelation.

$$r_s = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

$$r_s = \frac{1 - 6(1 \cdot 5)}{(6((36-1))}$$

$$r_s = 1 - \frac{9}{210} = 1 - 0.04286 = 0.9571$$

Spearman's Rank correlation coefficient = 0.96
(strong positive co-relation).

Q7. Determine the rank correlation for following data with shows the marks obtained in 2 quizzes in mathematics.

Marks in 1st quiz 6 7 5 9 8 4 8 7 6 10

Marks in 2nd quiz 8 6 7 8 7 6 10 5 8 10

→ Step 1 Assign ranks -

X	Rank of x	y	Rank of y
4	1	6	2.5
5	2	7	5
6	3.5	8	7
6	3.5	8	7

X	R(x)	Y	(R(Y)
7	6.5	5	1
7	6.5	8	7
8	8.5	10	9.5
8	8.5	10	9.5
9	10	8	7
10	11	6	2.5

$$r_s = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

After calculating differences –

$r_s \approx 0.69$

Spearman's rank correlation = 0.69 (Moderate positive Correlation)

Q8 Prove Baye's theorem and solve the following problem. A factory with 2 machines past records.
Machine 1 = 20% of item
Machine 2 = 80 % of item.
Further is shows 6% of item by machine 1 are defective, 1 % of item of machine 2 are defective. If defective item is drawn at random what is the probability that it was produced by machine 1.

→ Baye's theorem and probability of Machine 1 producing defective item.

Step 1: Define given data.

- Machine 1 produces 20% of items
  $P(M_1) = 0.20$
- Machine 2 produces 80% of items.
  $(P(M_2)) = 0.80.$

- Defect rate of Machine 1 → $P(D/M_1) = 0.06.$
- Defect rate of Machine 2 → $P((D/M_2) = 0.01.$

We need to find $P(M_1/D)$ the probability that a defective item was produced by machine 1.

* Step 2: Apply Baye's theorem.

$$P(M_1/D) = \frac{P(D/M_1)\,P(M_1)}{P(D)}$$

First, compute $P(D)$:

$$P(D) = P(D/M_1)P(M_1) + P(D/M_2)\,P(M_2)$$
$$P(D) = (0.06 \times 0.20) + (0.01 \times 0.80)$$
$$P(D) = 0.012 + 0.008 = 0.02.$$

Now, $$P(M_1/D) = \frac{0.06 \times 0.20}{0.02} = \frac{0.012}{0.02} = 0.6$$

The probability that a randomly drawn defective item was produced by machine 1 is 0.6 or 60%.

Q9 Consider a class whose students have obtained followings marks out of 50. in mathematics. Calculate mean, median, mode, geometric mean, harmonic mean.

Mark: 35, 40, 45, 49, 34, 47, 39, 25, 19, 35, 28, 48.

→ Step 1: Mean = $\frac{\Sigma x}{n} = \frac{442}{12} = \boxed{36.83}$.

Step 2: Median

Ordered data ⇒ 19, 25, 28, 34, 35, 35, 39, 40, 45, 47, 48, 49.

Median = $\frac{35+39}{2} = \boxed{37}$.

Step 3: Mode.

Most frequent value = $\boxed{35}$.

Step 4: Geometric mean.

$GM = (\Pi x_i)^{1/n}$

$\boxed{GM = 35.86}$

Step 5: Harmonic mean.

$HM = \frac{n}{\Sigma 1/x_i}$

$\boxed{HM = 34.67}$

→

Q10. A survey of CPA, 10 years ago average salary 74,914. An accounting researcher would like to know whether over the year's average salary is increased or not. 112 CPAs mean salary = 78695.

Standard deviation = 14530.

$H_0 : \mu = 74914$ against $h1 : \mu > 74914$

$Z_c = 78695 - 74914 / (14530/112)$

→ Given -

Past mean = 74914

Sample mean = 78695.

Standard deviation = 14530.

Sample size = 112

Hypothesis : $H_0 : \mu = 74914$

$H_A : \mu > 74914$

Step 1 : Compute Z statistic.

$$Z = \frac{\bar{X} - \mu_0}{6/\sqrt{n}} = \frac{78695 - 74914}{14530/\sqrt{112}}$$

$$= \frac{3781}{1372.36} = 2.75$$

Step 2 : Compare with critical value -

For $\alpha = 0.05$, Z-value = 1.645.

Since $2.75 > 1.645$ we reject H_0.

Evidence suggests that CPA Salary has increased over the years.

* * * * *