Assalam o Alaikum!
Aaj hum Data Science ka pehla sabaq shuru kar rahe hain.

Sabse pehle yeh samajhte hain ke **Data Science hoti kya hai**.
Is video mein hum yeh dekhenge:

- Data Science kya cheez hai
- AI aur Machine Learning se yeh kaisay alag hai
- Is field mein kaam karne ke liye kon kon se skills chahiye

Data Science ek aisi field hai jismein aap ko:

- Thora math ka knowledge (statistics)
- Thora computer ka kaam (programming)
- Aur jis area mein aap kaam kar rahe hain, uske bare mein knowledge hona chahiye

Yeh sab milke Data Science banti hai.

Log aksar confuse hote hain ke Data Science aur Machine Learning ya Deep Learning aik jaisi cheezein hain.
Lekin yeh alag alag hain.

Machine Learning ka kaam hota hai sirf aise model bananay jo data se kuch seekh saken.
Lekin **Data Science** ka kaam zyada bara hota hai — yeh real duniya ke problems solve karti hai, shuru se le kar end tak.

Data Science ka process kuch is tarah hota hai:

1. **Masla samajhna** — pehle yeh sochte hain ke problem kya hai aur humein kya result chahiye
2. **Data ikatha karna** — yeh dekhna ke data kahan se milega
3. **Data saaf karna aur samajhna** — jo data mila usse theek tarah tayyar karna
4. **Model banana** — koi formula ya system tayyar karna jo data ko samjhe
5. **Result dikhana** — apne kaam ka result simple logon ko samjhana
6. **Communication aur Visualization** — asaan visuals (graphs, charts) se explain karna ke kya seekha

---

Aam log jo technical nahi hote, jaise CEO ya manager, unko samjhana hota hai ke data kya keh raha hai.

Iske liye graphs, charts aur asaan zubaan ka use hota hai.

Ab baat karte hain ke **Data Science seekhne ke liye kya cheezein zaroori hain**:

1. Thora math ka ilm (Statistics)
2. Machine Learning ka basic idea
3. Computer language jese Python ka basic use
4. Graphs aur charts bananay ka tareeqa
5. Aur sabse important: **apni baat samjhana aana chahiye** — taake apna analysis managers ko asaani se samjha sako

---

Umeed hai ab aapko ek **basic idea** mil gaya hoga ke Data Science kya hoti hai aur ismein kya kya seekhna hota hai.

Overview of Data Science Process

Data Science Kya Hai?

Data Science ek aisi field hai jisme hum **data ka istemal karke masail ka hal nikalte hain**.
Isme kuch important cheezein hoti hain jaise:

- Thora math ka ilm (jaise numbers ka average lena)
- Thora computer ka kaam (jese coding)
- Thora knowledge us area ka jahan aap kaam kar rahe ho (jaise health, shopping, ya property)

Data Science Ka Kaam Kis Tarah Hota Hai?

Yeh kaam **kuch important stages** mein hota hai. Har step ka apna maqsad hota hai. Neeche unka asaan tareeqa se zikr hai:

---

1. Maslay Ko Samajhna (Problem Formulation)

Sabse pehle yeh sochte hain ke:

- Masla kya hai?
- Kya cheez pata karni hai?
- Kaun se data ki zarurat hai?

- Kya result chahiye?

Example: Agar hum ghar ki price ka andaza lagana chahtay hain to input ho sakta hai: bedrooms, area, location.
Aur output hoga: price ka andaza.

---

2. Data Ikatha Karna (Data Acquisition)

Jab masla samajh ajaye, to phir data dhoondhna hota hai.
Data mil sakta hai:

- Online websites se (jaise Kaggle)
- University ya govt. ki sites se
- Apne company ke records se

Yeh data maslay ko solve karne mein madad karta hai.

---

3. Data Tayyar Karna (Data Preparation)

Is stage mein do kaam hote hain:

a) Data ko samajhna (EDA)

- Yeh dekhna ke data kis type ka hai
- Kya numbers theek hain ya kuch galat hai?
- Kya kuch cheezein missing hain?

b) Data saaf aur seedha karna (Pre-processing)

- Galtiyan theek karna
- Missing values ko fill karna
- Agar zarurat ho to data ko scale karna (jese sab ko 0 se 1 ke darmiyan lana)

---

4. Data Ka Analysis Karna (Data Analysis)

Ab hum models banate hain jo data ko samajh kar result nikalte hain.
Steps yeh hote hain:

- **Sahi tareeqa chunna:** Jaise agar price predict karni hai to regression use karte hain.
- **Algorithms se model banana:** Jaise simple regression ya tree-based models
- **Check karna ke kaunsa model sabse behtar chal raha hai**

- **Akhir mein model ko deploy karna:** Yani model ko aise bana dena ke koi aur bhi use kar sake, jaise website ya app pe

---

5. Natijay Samjhana aur Dikhana (Communication & Visualization)

Model banane ke baad uska natija dikhana hota hai un logon ko jo technical nahi hote, jaise manager ya CEO.
Iske liye hum use karte hain:

- Graphs (bar chart, pie chart, scatter plot waghera)
- Asaan visuals jo data ko simple bana dein

Tools jese:

- Python libraries (Matplotlib, Seaborn)
- Power BI ya Tableau

Sabse zaroori cheez: Achi **baat samjhanay wali skill** — taake non-technical log bhi decision le saken.

Aakhri Baat

Data Science aik **complete process** hai jo aapko maslay ko samajhne se le kar result dikhane tak le jaata hai.

Har step important hai:

- Masla samajhna
- Data lana
- Data ko saaf karna
- Model banana
- Aur akhir mein asaani se samjhana

Jab aap yeh sab achi tarah seekh lein to aap kisi bhi mushkil maslay ko data ki madad se solve kar sakte hain.

Lecture 2: Understanding Types of Attributes Data -

1 Data Object aur Uske Features

Data Science mein hum aise data ke sath kaam karte hain jo table ki shakal mein hota hai.

- **Rows** (satrain): Har row ek shakhs, cheez ya record ko dikhati hai.

- **Columns** (ustoon): Har column us shakhs ya cheez ke kisi aik pehlu (feature) ko dikhata hai.

Example: Agar logon ka data ho to:

- Har row ek alag banda hoga
- Har column jaise height, weight, job, hair color waghera ko dikhata hoga

---

2 Attributes (Data ke Types)

Data ke features alag alag qisam ke ho sakte hain. Neeche unki asaan samajh hai:

➤ Nominal Attributes

- Yeh woh hotay hain jo sirf naam ya label hotay hain
- Jaise: hair color — black, brown, red
- Inka koi order nahi hota, aur in pe math ka koi formula apply nahi hota (jaise add karna)

➤ Binary Attributes

- Sirf do options hotay hain
- Jaise: male/female, smoker/non-smoker
- Coding mein inhein aksar 0 aur 1 se dikhaya jata hai
- Kabhi dono values barabar important hoti hain (jaise gender), aur kabhi ek zyada important hoti hai (jaise COVID positive)

---

3 Data Ka Tayyari Ka Amal (Data Preparation)

Data analysis shuru karne se pehle humein data ko samajhna aur tayyar karna parta hai. Iska 2 hisson mein kaam hota hai:

➤ Data Samajhna (EDA)

- Yeh dekhte hain ke data kaisa hai
- Kya kuch values missing hain?
- Kya koi value ajeeb ya galat lag rahi hai?
- Kya alag columns ka aapas mein koi link hai?

➤ Data Saaf Karna (Cleaning)

- Galat ya missing data ko theek karte hain
- Kuch values ko scale ya normal shape mein le aate hain
- Yeh sab karne ke baad data model ke liye tayyar hota hai

4️⃣ Attributes ko Samajhna

Data ke columns ko hum **attributes**, **features**, ya **variables** bhi kehte hain — ye teenon lafz aik hi cheez hain.

Example:
Agar aik shakhs ka record ho to:

- Naam: Ahmed
- Height: 5'9"
- Weight: 150
- Hair Color: Brown

Yeh sab us shakhs ke features ya attributes hain.
Aur puri row, yani yeh sab milke, **ek data object** kehlata hai — ek banda ya cheez.

Aakhri Baat (Final Thoughts)

Data Science mein kaam karne ke liye aapko kuch zaroori cheezon ki samajh honi chahiye:

- Samajhna ke data ka structure kya hai
- Har attribute ka kya matlab hai
- Data ko saaf aur sahi form mein laana
- Aur akhir mein analysis ya model ke liye use karna

Chahe attributes nominal hoon ya binary, unka role important hota hai.
Data ko samajhna, tayyar karna, aur usse sahi tareeqe se use karna — yeh sab **Data Science ke bunyadi steps** hain.

Agar yeh sab steps sahi se samajh liyein to aap kisi bhi data se **useful baatein nikaal** sakte hain.

Types of Attributes

1️⃣ Nominal Attributes (Naam Wale Features)

Kya hota hai:
Yeh aise features hote hain jo sirf naam ya label batate hain – inka koi order nahi hota.

Example:

- Hair color: Black, Brown, Red, White

Khaas baat:

- Inmein koi choti ya bari value nahi hoti
- Inko add ya subtract nahi kar sakte

- Sirf yeh dekh sakte hain ke kaunsi value kitni dafa aayi hai
  (jaise: kitne logon ke baal black hain)

2️⃣ Binary Attributes (Sirf 2 Options Wale Features)

Kya hota hai:
Aise features jismein sirf do hi values hoti hain

Example:

- Gender: Male ya Female
- COVID: Positive ya Negative
- Smoking: Smoker ya Non-smoker

Khaas baat:

- Sirf 2 choices hoti hain
- Kabhi dono barabar important hoti hain
- Kabhi aik zyada important hoti hai (jaise COVID positive)

Kya kar sakte hain:

- Count kar sakte hain ke kaunsi value kitni dafa aayi hai

3️⃣ Ordinal Attributes (Order Wale Features)

Kya hota hai:
Aise features jinmein values ka **order ya level** hota hai, lekin exact difference nahi pata hota

Example:

- Education Level: Junior, Assistant Professor, Associate Professor, Professor

Khaas baat:

- Values ka order hota hai (jaise choti se badi taraf)
- Lekin har level ka difference barabar nahi hota
- Jaise: Junior aur Assistant mein kitna fark hai — wo fix nahi hota

Kya kar sakte hain:

- Sabse zyada aayi hui value dekh sakte hain (mode)
- Beech wali value dekh sakte hain (median)

- Numbers ko rangon mein badal sakte hain (jaise temperature: Low, Medium, High)

4️⃣ Numeric Attributes (Numbers Wale Features)

Yeh wo features hain jo **numbers** mein hote hain — in pe math ke formulas lagte hain.
Yeh do qisam ke hote hain:

➤ Interval-Scaled

Example: Temperature (jaise: 20°C, 30°C)
Khaas baat:

- Zero ka matlab "kuch bhi nahi" nahi hota
- Numbers ke darmiyan gap barabar hota hai
- Add aur subtract karna meaningful hota hai

➤ Ratio-Scaled

Example: Height, Weight, Experience
Khaas baat:

- Zero ka matlab hota hai "bilkul nahi"
- 2 meter height, 1 meter se double hoti hai
- Sab math ke formulas apply hote hain: add, minus, multiply, divide

---

5️⃣ Discrete aur Continuous Attributes

➤ Discrete:

- Sirf **poore numbers** hote hain
- **Example:** Number of children (1, 2, 3…)

➤ Continuous:

- **Koi bhi value** ho sakti hai, decimal ke saath bhi
- **Example:** Height (5.6), Weight (60.3)

Aakhri Baat

Data science mein kaam karte waqt yeh samajhna bohot zaroori hota hai ke:

- Kis type ka data hai
- Har type ke data ke sath kaise kaam karna hai
- Kya us pe math ke formulas lagte hain ya nahi
- Kya order hai ya sirf naam hai

Jab yeh sab cheezein samajh aa jayein to:

- Data ko saaf karna
- Us pe analysis karna
- Aur model banana — sab kaam asaan ho jaata hai

Lecture 4: Understanding Statistical description in Data Science

Statistical Description kyun zaroori hoti hai?

Jab humein yeh samajhna ho ke kisi column ya feature ke andar values kis tarah banti hain ya bikhar gayi hain, to hum statistics ka sahara lete hain. Yeh mushkil nahi hota — asal mein yeh humein yeh batata hai:

- Data kis shape mein hai
- Values kis had tak phaili hui hain
- Aam ya beech wali value kya hai
- Data mein koi extra ya random cheezein to nahi (noise ya bias)

Yeh sab cheezein humein data ko samajhne, saaf karne, aur analysis ke liye tayar karne mein madad karti hain.

🔁 Data ke Distribution ke Types

1️⃣ Symmetric Distribution

- Data dono sides (left aur right) se barabar hota hai
- Beech mein ek line hoti hai jo average value ko dikhati hai
- Aisa data balanced hota hai — jise hum kehte hain "normal distribution"

2️⃣ Skewed Distribution

- Jab data barabar nahi hota — ek side zyada hoti hai

Types:

- **Positive Skew:**
  Data zyada left side par hota hai, aur right side slow slow kam hota hai
  Example: Aam log kam kamaate hain, lekin kuch log bohot zyada kamaate hain
- **Negative Skew:**
  Data pehle slow barhta hai, phir right side par achanak gir jaata hai
  Aisa data bhi common hota hai real life mein

📐 Statistics ke 3 Basic Measures

1️⃣ Central Tendency (Data ka beech ya aam point)

Yeh batata hai ke data ka "center" ya aam tor par value kya hai:

- **Mean:** Sab values ka average
- **Median:** Beech ki value
- **Mode:** Sabse zyada aayi hui value
- **Trimmed Mean:** Outliers hata ke average
- **Midrange:** Sabse choti aur badi value ka average

2️⃣ Dispersion (Data kitna phaila hua hai)

Yeh batata hai ke values average ke around kitni door ya paas hain:

- Agar values paas paas hain → data tight hai
- Agar values door door hain → data spread out hai

Yeh humein batata hai ke data mein kitni consistency ya variation hai

---

3️⃣ Similarity ya Proximity (Features aapas mein kitne milte hain)

Yeh dekha jaata hai ke 2 ya zyada columns ya features aapas mein kitne similar hain:

- Agar 2 features bohot milte hain, to aik ko hata bhi sakte hain
- Analysis simple aur fast ban jata hai

🔊 Noise aur Bias samajhna

📉 Noise (Random galti ya disturbance)

Noise ka matlab hai aisi random cheezein jo data mein asar daalti hain:

Examples:

- Weight machine har baar alag number dikha rahi ho
- Quiz de rahe ho lekin kabhi noise ki wajah se score kam aaya

Khaas baat:

- Random hoti hai
- Har dafa alag hoti hai
- Predict nahi kar sakte

📏 Bias (Hamisha aik taraf ka error)

Bias matlab hamesha aik tarah ki galti hona:

Example:

- Height measure karne wala tool har kisi ki height ko 8 inch zyada dikha raha ho

Khaas baat:

- Har dafa wahi galti karta hai
- Aik hi direction mein hoti hai
- Data ko misleading bana deta hai

Noise vs Bias

Noise	Bias
Random aur alag alag	Fix aur same
Har dafa different	Har dafa same
Predict nahi hota	Predictable error hota hai

Outliers kya hote hain?

Outliers wo values hoti hain jo baaqi data se bohot alag hoti hain.
Example: Agar sab log 20–30 saal ke hain, lekin ek value 80 saal ho to wo outlier hoga.

Practical Faida

Jab hum basic statistics samajh lete hain, to:

- Noise aur bias ko samajh sakte hain
- Outliers dhoondh sakte hain
- Data ka shape aur spread samajh sakte hain
- Aam value maloom kar sakte hain
- Features ke darmiyan relationship samajh aata hai
- Predictive models behtar ban sakte hain

Aakhri Baat

Data ko analyse karne ka pehla qadam hai uska basic statistical description:

- Beech wali value kya hai (central tendency)
- Data kitna phaila hua hai (dispersion)
- Features aapas mein kitne related hain (similarity)

Yeh sab aapko data samajhne aur us par sahi analysis karne mein madad deti hain.

Yeh foundation banata hai Data Science ke aage ke topics ko samajhne ke liye.

Lecture 5: Measure of Central Tendency

Central Tendency kya hoti hai?

Central Tendency ka matlab hai data ka wo number jo beech mein hota hai:

- 50% data is number se neeche hota hai
- 50% data is number se upar hota hai

Yani yeh wo value hoti hai jo data ka **balance point** hoti hai. Jaise jab hum logon ki height ya salary dekhte hain, to zyadatar values ek beech wale number ke ird gird hoti hain.

Central Tendency ke common tareeqe:

1. **Mean** (Average)
2. **Median** (Beech wala number)
3. **Mode** (Sabse zyada baar aane wali value)
4. **Trimmed Mean** (Outlier hata ke average)
5. **Mid-Range** (Sabse choti aur badi value ka average)

Mean (Average)

Kya hai?
Sab values ko jod kar, total number se divide kar dena.

Example:
Agar 4 log 100,000 kama rahe hain, aur aik banda 20 lakh kama raha hai,
to average salary banegi lag bhag 4.8 lakh
(halanke zyadatar log 1 lakh hi kama rahe hain)

Masla:
Outliers (bohot zyada ya bohot kam value) isay disturb kar dete hain

---

Trimmed Mean

Kya hai?
Aise average jo extreme values hata ke nikala jaye.

Example:
Agar kisi ka salary data mein aik value bohot high hai (jaise 2 million), to usay hata kar baqi values ka average nikalte hain.

Tip:
Zyada se zyada 2% data trim karna chahiye

Median

Kya hai?
Sorted list mein beech wala number

Example:

- Agar values: [2, 4, 6, 8, 10] → Median = 6
- Agar values: [2, 4, 6, 8] → Median = (4 + 6) / 2 = 5

Faida:
Outliers ka asar nahi hota
Nuksan:
Data ko sort karna padta hai, zyada data ho to slow hota hai

Mode

Kya hai?
Jo value sabse zyada baar repeat ho

Example:
10 log ki height 5'8" hai, aur 9 log ki 6'1", to dono modes hain
(Multimodal data)

Faida:
Non-numeric data (jaise color, city, etc.) ke liye bhi kaam karta hai

---

Mid-Range

Kya hai?
(Minimum + Maximum) / 2
Sirf ek jaldi andaza lagane ke liye use hota hai
High accuracy ke liye nahi

Central Tendency se kya samajh aata hai?

Yeh measures humein yeh batate hain ke data ka shape kaisa hai:

Agar:

- Mean ≈ Median ≈ Mode → Symmetric data (balanced)
- Mode < Median < Mean → Positive Skew (right side heavy)
- Mean < Median < Mode → Negative Skew (left side heavy)

Example:
Agar aap ke paas 10,000 rows ka data hai aur aap graph nahi bana sakte, to:

- Mean, Median, Mode nikaalo
- Compare karo → data skewed hai ya balanced

Final Baat

Central Tendency data science ka pehla aur basic qadam hai:

Is se aap:

- Beech wali value samajh sakte ho
- Data ka shape aur balance dekh sakte ho
- Preprocessing (data ko tayar karna) asaani se kar sakte ho

Lecture 6: Measure of dispersion

Dispersion ka matlab kya hota hai?

Data analysis mein **dispersion** ka matlab hai ke data kitna **faila hua** ya **gathha hua** hai apne beech ke number ke aas paas. Is se humein yeh pata chalta hai ke:

- Data consistent hai ya nahi
- Sab values ek jaisi hain ya alag alag
- Decision lena aur prediction karna kitna asaan hoga

Agar aapko data ka spread samajh aa jaye to aap zyada **ba-himmat aur sahi faislay** le sakte hain.

Basic Concepts: Central Tendency aur Dispersion

- **Central Tendency**: Data ka beech ka number — jaise mean, median, ya mode
- **Dispersion**: Data us beech ke number ke aas paas kitna faila hua hai

---

Quartiles aur IQR kya hote hain?

- **Quartiles**: Data ko 4 barabar hisson mein divide karne wale points
  - **Q1**: 25% data is se neeche hota hai
  - **Q2**: 50% data ka beech (median)
  - **Q3**: 75% data is se neeche hota hai
- **IQR (Interquartile Range)**:
  Q3 - Q1 → beech ke 50% data ka spread
  - **Chhota IQR** → data gathha hua hai
  - **Bara IQR** → data faila hua hai

5-Number Summary

Yeh 5 numbers data ka full snapshot dete hain:

1. **Minimum value** (sabse chhoti value)
2. **Q1**
3. **Median (Q2)**
4. **Q3**
5. **Maximum value** (sabse badi value)

Data Symmetry aur Skewness

- **Symmetric Data**:
  Q1, Q2, Q3 ka distance barabar hota hai
- **Skewed Data**:
  Agar left ya right side zyada faili hui ho

Box and Whisker Plot kya hai?

Yeh aik visual graph hota hai jo upar wale 5-number summary ko dikhata hai:

- **Box** → Q1 se Q3 tak ka data (IQR)
- **Line in box** → Median
- **Whiskers** → Box ke bahar ka range
- **Dots** → Outliers (jo data se bohot zyada alag hain)

Outliers kaise nikaalte hain?

- Lower: Q1 – 1.5 × IQR
- Upper: Q3 + 1.5 × IQR

Standard Deviation aur Variance kya hoti hai?

Yeh dono cheezein **tab use hoti hain jab mean** ko central value maana jaye:

- **Standard Deviation**:
  Batata hai ke har value mean se kitni door hai
  (jitni badi SD, utna zyada spread)
- **Variance**:
  Har value ka mean se difference square karke average nikaalte hain

---

Kab kaunsa use karein?

Situation	Tool
Median center ho	Quartiles, IQR, Box Plot
Mean center ho	Standard Deviation, Variance

Practical Example

Aap ke paas games ki yearly sales ka data hai:

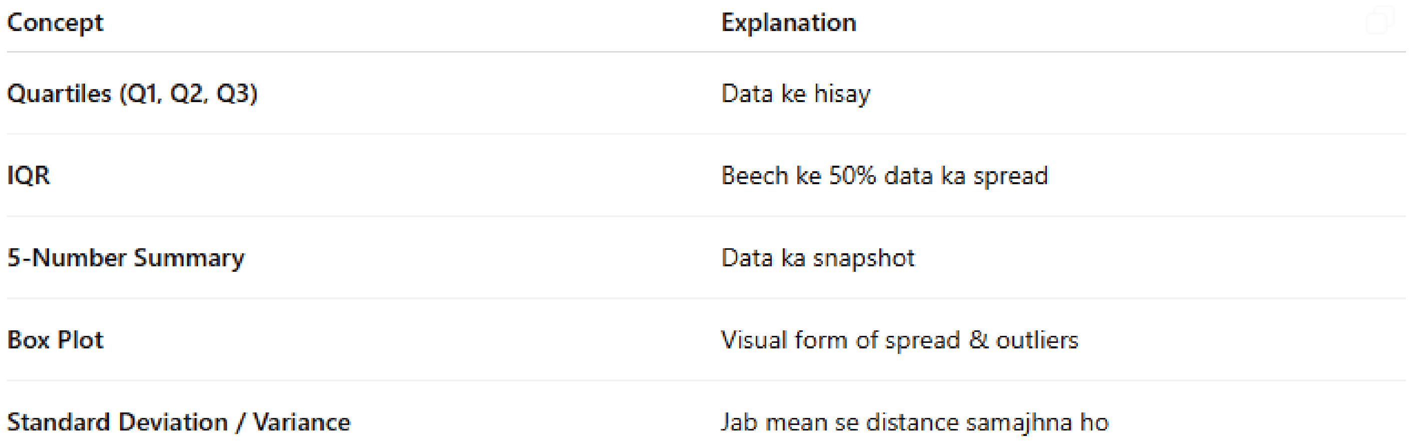
- **IQR** batayega ke beech ke 50% games ka sales kitna vary karta hai
- **Box Plot** dikhayega koi extreme value (outlier) hai ya nahi
- Agar koi game bohot zyada bika ho, to outlier hoga

Real Life Mein:
Outliers nikaal kar aap marketing ya quality problems samajh sakte ho.

Final Baat

Dispersion aapko yeh samajhne mein madad karta hai:

- Data gathha hua hai ya spread hai
- Outliers kahan hain
- Data symmetric hai ya nahi
- Kis tarah ka model ya tool use karna chahiye



Concept	Explanation
Quartiles (Q1, Q2, Q3)	Data ke hisay
IQR	Beech ke 50% data ka spread
5-Number Summary	Data ka snapshot
Box Plot	Visual form of spread & outliers
Standard Deviation / Variance	Jab mean se distance samajhna ho

Lecture 7: Introduction to Proximity & Similarity

Proximity ya Similarity Measures kya hotay hain?

Data Science mein jab hum kisi dataset ke **do features** ya **attributes** ke darmiyan **taluk** ya **similarity** samajhna chahtay hain, to hum **proximity** ya **similarity measures** use kartay hain.

Yeh measures humein yeh batatay hain:

- Do cheezain kitni milti julti hain
- Kya unka direction ya trend ek jaisa hai
- Kya wo dono sath barhti ya kam hoti hain

---

1. Dot Product

Kya hota hai?
Dot product aik mathematical tareeqa hai jo batata hai ke do vectors (features) ka **direction kitna same hai**.

Simple Example:
Do log aik gari ko dhakka day rahay hain:

- Aik agay se aur aik peechay se → koi move nahi → dot product = 0
- Dono aik hi side se dhakka den to gari chalti hai → dot product positive

Downside:
Dot product mein agar values ka size change ho jaye (jaise unit ya measurement change ho), to result bhi change ho jata hai. Yani yeh **magnitude** pe depend karta hai.

---

2. Cosine Similarity

Kya hota hai?
Cosine similarity direction ko measure karta hai **baghair magnitude ke**. Yani values ka size matter nahi karta — sirf yeh dekhta hai ke direction match karta hai ya nahi.

Range:

- 1 → Bilkul same direction
- 0 → Bilkul different (orthogonal)
- -1 → Opposite direction

Example:
Temperature aur ice cream ki sales — dono barhtay hain sath sath → similarity high (close to 1)

Downside:
Yeh bas direction batata hai, lekin **taalluq ki taqat (strength)** nahi batata.

---

3. Covariance

Kya hota hai?
Covariance dekhta hai ke do variables ek sath **barhtay ya ghat'tay** hain ya nahi.

- Agar dono sath barhtay hain → **Positive covariance**
- Aik barhta aur doosra ghat'ta hai → **Negative covariance**

Example:
Temperature barhne se ice cream sales bhi barhti hain → covariance positive

Downside:
Yeh **sirf direction** batata hai, lekin **kitna taqatwar relation hai**, yeh nahi batata.

📈 4. Correlation

Kya hota hai?
Correlation **covariance ka upgraded version** hai jo **direction ke sath sath strength bhi batata hai**.

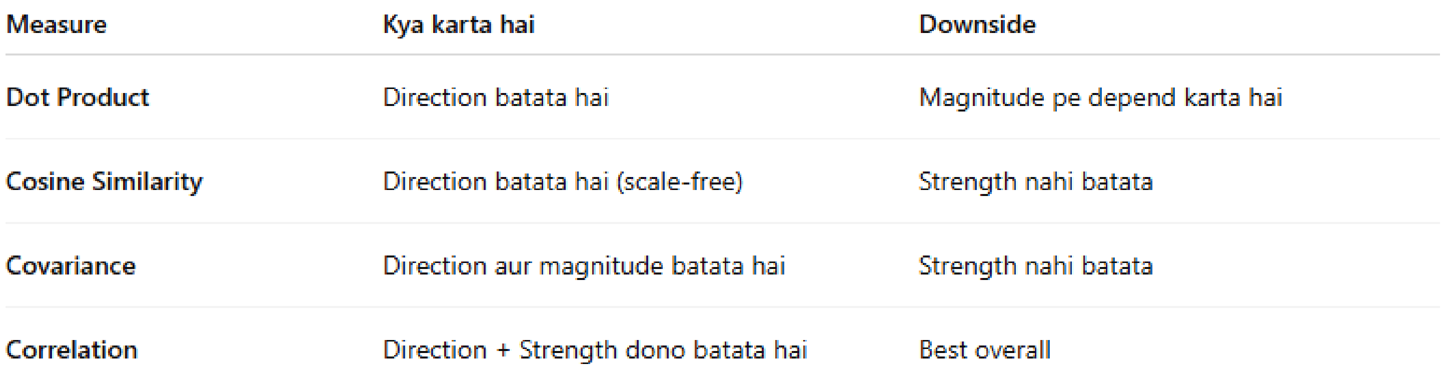
Range:

- 1 → Perfect positive relation
- 0 → Koi relation nahi
- -1 → Perfect negative relation

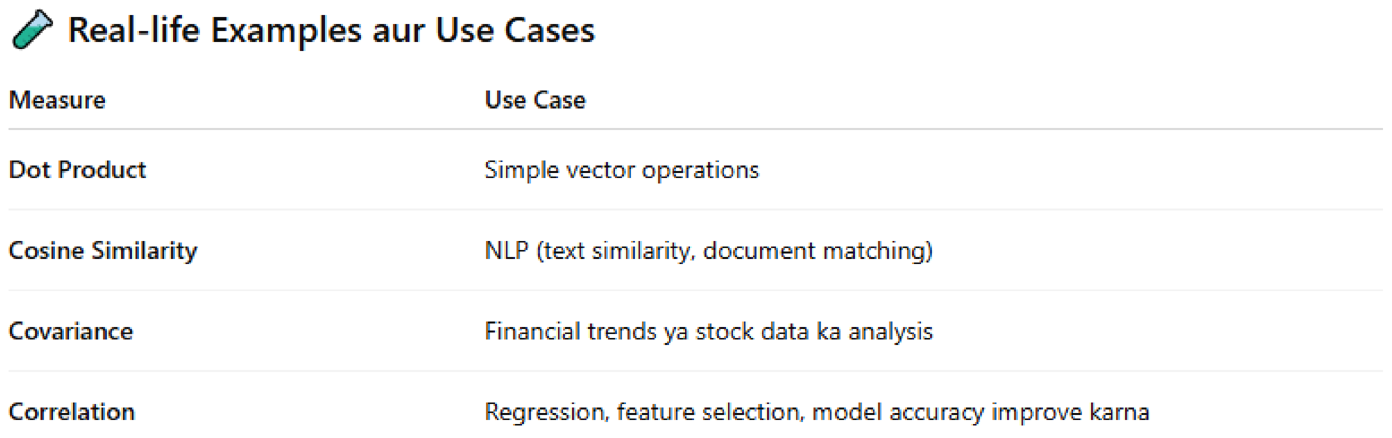
Use Cases:

- Features ka taluk samajhne ke liye
- Feature selection aur dimensionality reduction mein use hota hai

Example:
Agar temperature barhne par sales consistently barhti hain → strong positive correlation



Measure	Kya karta hai	Downside
Dot Product	Direction batata hai	Magnitude pe depend karta hai
Cosine Similarity	Direction batata hai (scale-free)	Strength nahi batata
Covariance	Direction aur magnitude batata hai	Strength nahi batata
Correlation	Direction + Strength dono batata hai	Best overall



🧪 Real-life Examples aur Use Cases

Measure	Use Case
Dot Product	Simple vector operations
Cosine Similarity	NLP (text similarity, document matching)
Covariance	Financial trends ya stock data ka analysis
Correlation	Regression, feature selection, model accuracy improve karna

Final Baat

In measures ko samajhna bohot zaroori hai data science mein:

- Aap samajh sakte hain ke kaunsi cheez kis cheez se kitni related hai
- Data ko better analyze aur clean kar sakte hain
- Model training mein better decisions le sakte hain

Sabse best aur commonly used measure — Correlation hai, kyun ke yeh direction bhi batata hai aur taqat bhi.

Lecture 8: Data Quality and Data Preprocessing

1. Data Quality kya hoti hai?

Data Quality ka matlab hota hai ke aapke paas jo data hai, wo **sahi**, **clean**, aur **reliable** ho — taake us par machine learning model train kiya ja sake jo achi prediction day.

Achi quality ke data ki khasoosiyat:

- **Sahi ho (Accuracy)** – Ghalat values na ho
- **Consistent ho** – Confusing ya contradict karne wali values na ho
- **Duplicate na ho** – Ek hi data baar baar na likha ho
- **Missing values kam hon** – Jitna ho sake complete data ho
- **Update ho (Timely)** – Purana ya outdated data na ho

Example:
Agar kisi ka address 10 saal purana likha hai, to wo data **timely nahi** hai.
Agar kisi ki age 604 likhi hai, to wo **consistency ka masla** hai.

⚠ 2. Real Life mein Data ke Maslay

Jab data real world se collect hota hai, to aksar usmein ye problems hoti hain:

- ❌ **Missing values** – Kahi columns blank hote hain
- ❌ **Duplicate records** – Ek hi banda ka data 2 ya 3 dafa likha hota hai
- ❌ **Wrong values** – Jaise age 800, ya weight -50
- ❌ **Fazool features** – Jo analysis mein kaam nahi aatay

Garbage In, Garbage Out:
Agar data kharab hoga, to model bhi kharab result day ga. Is liye **data ko pehlay saaf karna zaroori hai**.

---

3. Data Pre-processing – Data ko Behtar Banana

Pre-processing ka matlab hai **data ko model ke laayak banana**. Yeh kaam bar bar karna parta hai aur kuch key steps hain:

✅ **Data Cleaning:**

- Duplicate data hatao
- Fazool columns hatao
- Missing data fill karo ya remove karo

🔗 **Data Integration:**

- Alag alag sources ka data ek jagah combine karo
- Jaise lab results, DNA reports, aur patient info — sab ek file mein lao

📉 **Data Reduction:**

- Same type ke features merge karo
- Jo columns repeat ya irrelevant hon, unko hatao
- Techniques jaise PCA use karo taake columns kam ho jayein lekin info barkarar rahe

🔁 **Data Transformation:**

- Features ko same scale par lao (0 se 1)
- Agar aik column ka range 1-100 aur doosre ka 1-5 crore ho, to bara number model ko bias kar deta hai
- Normalize karo using Min-Max, Z-score, etc.

Example:
Age (1–100) aur income (1 lac – 5 crore) dono ko 0–1 scale par le ao — taake dono ka equal effect ho model pe.

🔄 4. Pre-processing aik baar ka kaam nahi

Data ko clean karke jab aap alag sources ka data combine karte ho, to naye issues samne aate hain. Is liye:

- Kabhi cleaning dobara karni padti hai
- Kabhi transformation ke baad feature reduction zaroori ho jata hai
- Har step baar baar revise karna parta hai

Yeh process **iterative** hoti hai — matlab **repeat hoti hai jab tak data best form mein na ho.**

🧠 5. Akhri Baat (Final Thoughts)

Agar data quality achi hogi to:

- Model accurate hoga
- Predictions reliable hongi
- Business problems ka solution behtar milega

Important steps:

- Cleaning
- Integration
- Reduction
- Transformation

Yeh sab milke **machine learning ke liye strong base tayar kartay hain.**
Agar data ghalat hoga to model bhi ghalat hi seekhay ga.

Lecture 9: Data Cleaning

Pichlay lecture mein hum ne **Data Pre-processing** ka intro discuss kiya tha — ke data ko model bananay se pehlay hum kis tarah tayar karte hain.

Hum ne 4 important kaam (steps) bataye thay:

- **Data Cleaning** (data ko saaf karna)
- **Data Integration** (alag alag sources ka data milaana)
- **Data Reduction** (fazool ya zyada features ko kam karna)
- **Data Transformation** (data ko model ke liye readable banana)

Phir hum ne ye bhi samjha tha ke **Data ki Quality** kya hoti hai — aur agar data ki quality poor ho, to model bhi **galat kaam karega**.
Jaise:

- Missing values ho
- Ghalat ya inconsistent values ho
- Noisy ya old (buhat purana) data ho

To Data Pre-processing ka maqsad hota hai ke data itna behtar ho jaye ke jab model is pe train ho, to wo **sahi insights aur patterns seekh sake.**

Aaj ka topic: Data Cleaning

Data cleaning mein aap thora **detective** ka kaam karte ho. Pehlay aapko yeh dekhna hota hai ke:

- Kahan kahan values missing hain?
- Kya koi data galat ya confuse karne wala hai?
- Kya koi row ya column purana ya irrelevant hai?

Phir aap un problems ko solve karne ke tareeqay use karte ho.

📊 Example: Gym Data

Sochiye aap ek **Gym** chala rahe ho. Aapke pass members ka data hai — jaise:

- Member ID
- Age
- Weight
- Profession
- Address
- Favourite Machine
- Kab join kiya
- City, etc.

Yeh example hum pura topic explain karne ke liye use karenge.

Missing Data ke 2 Scenarios

🟥 Scenario 1: Row-wise Missing Data

Kuch members aise hain jinka sirf **ID ya naam** hai — baaki sab kuch missing hai.
Yani 90% data unke baare mein missing hai.

- Aise rows ko **remove karna** behtar hota hai, warna model confuse ho jata hai.
- Agar sirf 5–10% rows aisi hain, to safely delete kar do — baaki 90% data se model achi training le lega.

Scenario 2: Column-wise Missing Data

Ab sochiye ke **"Favourite Machine"** ka column hai — lekin 90–95% logon ke liye value missing hai.

To aise column ka koi faida nahi — isay bhi **drop kar dena** chahiye.

Scenario 3: Thori thori missing values

Agar kisi attribute mein sirf **kuch values** missing hain (jaise 5–10%), to hum unko **fill** kar sakte hain.

Method 1: Constant value se fill karna

Example:
Agar age missing hai, to -1 likh do
Ya categorical column mein "None" likh do

Lekin dikkat yeh hai ke model in values ko **asli data** samajh sakta hai, aur galat pattern seekh sakta hai.

Method 2: Mean/Median/Majority se fill karna

Agar data symmetric hai (jaise bell shape), to **mean** use karo
Agar data skewed hai, to **median** use karo (kyunke wo extreme values se affect nahi hota)

Example:
Agar kisi ka **weight missing** hai, to sab members ka average weight nikal ke wahan likh do.

Thora advanced method

Agar aapke paas labels hain (jaise: Regular gym members vs Irregular), to aap unhi group ka average nikaal ke missing value fill kar sakte ho.

Example:
Agar koi **regular** member hai jiska weight missing hai, to regular members ka average weight usko assign karo.

Summary:

- Agar buhat zyada data missing ho (poori row ya column), to remove kar do
- Agar thori si values missing hoon, to fill kar sakte ho:
  - Constant value (e.g. -1, None)
  - Mean, Median ya Mode
  - Group-wise average

Yeh basic techniques hain **missing data** ko handle karne ke liye.
Lekin yaad rakho: har dataset alag hota hai — is liye thori **trial and error** bhi zaroori hoti hai.

Agar aap aur explore karna chahte ho, to ChatGPT ya kisi aur AI tool se aur bhi advanced methods seekh sakte ho.

Lecture 10: Resolving Inconsistent Values

Pehla masla humne solve kiya tha — **missing values** ka. Lekin doosra masla bhi tha — **inconsistency**.

Iska matlab hota hai ke data ek jaisa nahi likha gaya. Jaise gym ke data mein "City of Birth" ke andar **Lahore** likhne ke kai tareeqe hain. Kahin capital letters mein **LAHORE**, kahin **LHR**, kahin small letters mein **lahore**. Ye sab ek hi cheez hain lekin tareeqa alag alag hai — isay kehte hain **inconsistent data**.

Agar hum aise data ko model mein daalenge, to system har version ko alag cheez samjhega — jisse asli insight khatam ho jaayegi. Iska solution hai **standardization** — yani decide karna ke hum Lahore ko sirf **"lahore"** (small letters mein aur poori spelling ke saath) likhenge.

Iske liye ek chhoti si script likh lete hain jo har jagah jo bhi variation ho (jaise LHR, lahore, LAHORE), sabko **"lahore"** mein badal deta hai. Isi tarah Karachi ya Faisalabad ke saath bhi karte hain.

Dusra tareeqa hota hai **validation** — jisme hum dekhte hain ke ek record ke do values aapas mein match karti hain ya nahi.

Jaise ek member ki **Date of Birth** se calculate ki gayi **age** 30 saal banti hai, lekin record mein likha hua hai 45 saal. Yani mismatch hai. Phir hum decide karte hain ke hum kis value ko trust karenge — calculated age ya written age? Ye decide karne ke baad hum data ko clean kar lete hain.

Agla masla hota hai **duplicate data** ka — yani ek hi banda ka data system mein 2 dafa aa gaya. Ye kabhi kabhi hota hai jab form 2 dafa submit ho jaye ya kisi wajah se system mein ek hi cheez 2 dafa store ho jaaye.

Agar humne member ID unke **CNIC** pe banayi ho to duplicate pakarna easy hota hai — kyunke CNIC duplicate nahi hota. Lekin agar humne apne tareeqe se random ID di ho jaise "CYS13" waghera, to phir humein dekhna padega:

- Kya name same hai?
- Kya age, weight, aur dusri cheezen match karti hain?

Agar sab cheezen match karti hain to shayad wo duplicate ho sakta hai. Nahi to nahi.

Uske baad aata hai **noisy ya inaccurate data** — yani galat ya bekar data. Jaise kisi bande ki age likhi ho **270 saal** ya **573 saal** — jo clearly possible nahi.

Aise data ko **outlier** kehte hain. Hum decide karte hain ke gym ka member **18 se 60** saal ke darmiyan hoga. Is range ke bahar koi bhi age ho to hum usay galat ya suspicious maanenge.

Aise outliers ko ya to delete karte hain, ya unko "missing value" bana ke fill karne ki techniques lagate hain jaise humne missing data ke case mein kiya tha.

Aakhri aur important masla hota hai **stale data** — yani purana data jo update nahi hua. Jaise kisi member ka address 10 saal pehle ka ho. Agar hum us address pe email ya brochure bhejenge to shayad wo banda wahan na ho.

To hum check karte hain ke **last updated** kab hua tha? Agar 2 saal se zyada guzr chuke hain to samajh jaate hain ke wo data purana hai. Hum sirf un logon ko contact karte hain jinka data recent hai.

To dosto, is puri discussion mein humne yeh seekha:

- Missing values kaise handle karte hain
- Inconsistent data ko kaise standardize karte hain
- Duplicate records kaise identify karte hain
- Noisy ya outlier data ko kaise clean karte hain
- Stale data kaise detect aur manage karte hain

Ye sab cheezein milke data ko **accurate, saaf aur useful** banaati hain — jisse hum **real world problems** ko solve kar sakte hain.

Agle lecture mein hum **data integration** ke baare mein baat karenge.

Tab tak khuda hafiz aur duaon mein yaad rakhiye.
InshaAllah phir mulaqat hogi.