بسم الله الرحمن الرحيم

Data Science Cources

(1)

Introduction

Definition :-

- A Data Science is an interdisciplinary field that includes
  - → Statistics
  - → Artifical Intelligence
  - → Programming
  - → Domain Expertise

Question :- Differenciate B/w Machine learning, Deep learning and Data Science?

ML :-	VS	Data Science :-
→ Focuses on studing models and algorithms that learn from data		→ Works one steps from ML.
→ Just we interested in Model in ML/DL		→ Solves Real-world Problems.
		→ Follows an End-to-End Process from problem formulation to presentation

End-To-End Data Science Process:-

Data Science Cover the entire Journey from problem formulation insight divery.

- ★ Problem formulation
  Defining the question
- ★ Model Development
  Building Predictive models
- ★ Insight Communication
  Sharing Result
- ★ Data Acquisition
  Gathering Relevant Data
- ★ Model deployment
  Implementing in Production

◎ you can Represent all insight ...ayment to data Science Managment.

> "Data Science is End-to-End Discipline To deliver insight from with the help of Stat and ML Models."

Step 1:- (Role of domain Expertise)

★ Problem formulation :-

Exp:- House prices Predicting Application

★ Input features ★ Expected output

Loclion, Size, Age house price of house. (Selecting Appropride algorithms)

This insight of huse To Prediction Prices

Step 2:- ★ Data Acquisition (I/P)

Collect Relevant data for problems

data Sources

★ APIs	★ Data-base	★ Web-Scraping
→ Programmtic interface To Access data	→ Structed collection of information	→ Extracting data from web-sites

★ Public datasets :- → Open Source collection

Reporsity :- Kaggles, Many university.

Step 3:- Data Preparation (EDA)

→ Understanding and cleaning data Preprocessing.

✱ Cleaning data	✱ Transforming data
→ handing missing values and Removing outliers	→ Scaling, encoding categorical variables

✱ **Feature Engineering:-**

→ Creating new meaningful variables.

Step 4:- Data Analysis / Model Development:

✱ Selecting machine learning models "Evaluating" and "fine tuning" models Performance matrics

→ Collection Ratio → True-Positive Accuracy

→ finding all (Prediction(positive)

Step 5:- Model Development

Trained models accesible for real world use. **Methods:-**

✱ Web-application	✱ Embedded Systems.
User friendly interface	hardware implementation

✱ **APIs:-** Service Integration.

Step 6:- Communicating insight:-

→ Sharing these insight to (non-technical) Stakeholders (CEO, Managment) all descions maker.

★ data visualization	★ Storytelling
→ graphical represen- of your insight	→ Crafting with narratives of data

→ Business Reports:- formal - documentation

Question:- Which types of skill are required for Data Science?

Essential Skills:-

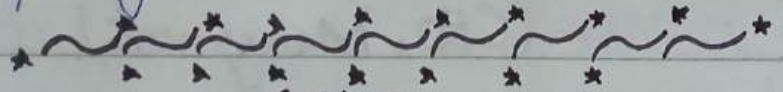
1 Statistics
- Understanding data distribution

2 Machine learning
- Predicted models

3 Python-Programming
- Implementing Solution

4 Data Visualization
- Creating graphs

5 Communication Skills
- Finding Clearty Exploring.



1step:-

(EDA):- An other method use in data understanding.

Exploring the data types.

Structed data / unStructed Data

if Structed types then we have seen which "types of" data."

Attributes: Normal values, ordinary types.

→ Data Types
- Numerical, category binary

→ Distribution of values
- how data is spread.

→ Missing values and outliers:-
- data quality issues.

→ Central tendancy Measure:-
- Mean, Mode, Median.

→ Dispersion Measure:-
- Variance, Stand deviation

→ Proximity Measure
Corellection Similarty.

(2nd Steps:- Data Preprocessing:

1. Handling Missing values
- imputation or removal

2. Recovering Inconsistance
- fixing errors

3. Data Transformation
- Scaling (0 to Range of 10) normalization
(A house which has prices is "50" Rs This is very inconsistance)

4. Data Reduction
- Feature Selection

Impact of Data Processing:-

- Clean Structed data ensurees effects model training. It helps:-
  - → Improve model Accuracy :- Batter Predictions
  - → Reduce Computational time :- Faster Processing
  - → Eliminate Misleading information :- (Give More reliable Result.)
- اس سے بچاری ڈیٹا اس فارم میں کیا جاتا ہے کہ کوئی بھی DL/ML, Statistics کے الگورتھم کو استعمال کر سکے۔

Next Stage :- Step 3

Data Analysis :-

- The Core Stage when models are "developed." Use different Techniques in ML/AI Tools

(1) Regression :-
- Predicted Continuose value

(2) Classification
- Categorizing data

(3) Clustering :-
- Identifying Groups

(4) Statistical Modeling :-
- Understanding Relationships

(5) Self-Supervised Learning

Exp:- like house-price prediction (Score) Using Regression Algorithms models

* Algorithms
- → It is Just Techniques like:- Linear Regression, DL (Interpretable model)
- → Polynomial Regression (non-Linear Relationships)
- → Random Forest Regressor (Decision Tree)

* Models
- → Algorithem to see data and learn data make models.
- → Data Knowledge.

A Pictures worths Thousand Words

Evaluating Models:-

→ It helps measure model Performance using evaluation metrics

MSE

→ Mean Squared Error

- Average Sequare diff b/w predicted and actual values

RMSE

→ Root Mean Squared Error

- Square root of MSE in same units as target

R^2 :- (R-Squared Value)

→ Proportion of variance explained by models.

Model Development:-

- It is very important part of data analysis. So User of shakeholder can used and share experience and exprirment.

انٹرنیٹ پر Server پر ڈال کے اور سوال کرتے۔

→ Web-Application

- Browers-based interface

→ Mobile Application

- On Smart phones and Tablets

→ APIs

- Application-programming interface

→ Embedded System

- hardware devices

Communication & Visualization

- Data driven insight must be represented in graph. Used Tools

Matplotlib	Seaborn	Power BI/Teblea
python plotting library	Static data visual	Business Intel gence Tools

Types:- → Bar charts → Line charts → Scatter plots → Pie charts

(Lectures No# 3+4)

Topic:- Data Science Processes
Understanding of Data - Types of Attributes.... (EDA)

- Analyzing data distribution
- Detecting outliers
- Measure similarty b/w data objects. (Relation).

Data - Types:-

- "Structure Data" which is easy to understand (like, Speadsheet, Tables)

Un-Structured Data:-

Speech Signals, images, text
Written Text.

IMPORTANT:-

- Unstanding different types
- of Attributes in data

What is an Attributes?

- Column i an Excel sheet representing a Property of data
  Alternate Names
  Attributes. from "data base"
  Feature. frome "Machine Learning"
  Variable. from "Math"

Example:-

- A data of Person:
  Name, Hight, weight, hair color.

Data Object:-

A Row in a data sets representing an individual entity

Example:
In a person, data sets, each row represents a different person (Irfan, Sheeraz, Mujtaba)

Simple Terms:-

- Simple in statistic
- Observation in data (Research anoylsis)
- Record in database
- Instance in "Machine Learning"
- Data Points in visualization

Types of Attributes:-

Four major types of Attributes:-

- Nominal Attributes → Binary Attrib
- Ordinal Attributes → Numeric " "

★ Nominal Varible

- nominal Attributes categorical data that Serve as Lables or Names without any quantitative value or natural "order"
- Example:- Colors of hair (black, white ...) blood groups (A, B, AB, O) or only countrys names Inherent order is not appear.

- ★ Possible Encoding
- can Represent Using Symbols or code (black = 0, white = 1)
- only one operation of math is Apply Frequency
- Meanin operation most frequently hair colors.

★ Binary Attributes:-

→ It is a Special types of nominal attributes. only
→ two values.

Examples:-

→ Gender : Male / Female
→ Covid : Positive (1) / Negative (0)
→ Smoking status : Smoker or non smoker

★ Sub-Types:-

→ Symmetric Binary Attributes
Both values are equally important
Example:- Genger (Male / Female)

→ Asymmetric Binary Attributes:-
Example:- Smoker (1 = Smoker, 0 = Non-smoker)
One value is more importan

★ Interchange able words (class, catorey, Lables)

★ Ordinal Attributes:-

◉ Attributes whose values posses a meaningful order or Ranking between "categories" allowing for comparative relationships.

IMPORANT Characteristics

◉ while we can determine if one value is Greater Than, other, the magnitude of differences b/w values cannot be precisely quantified.

Example:- Academic Grade A+ (ellent
→ A (very Good). B (Good) ....
Rands : Assistant Professor, Associate
→ Professor, full Professor.

Represent:- Code, Seimple not numerical
→ values Two features most frequent values is (modes) .: Between values (Median), Extract from numerical (Not perform arithmetic operation)

is a way To Convert numeric attribute to ordinal attributes

Continous value : folting point values: 1.23
Discrete value : whole numbers : 2, 4, 6

★ Discretization (convert) & ordinal Variables :-

- Convert a numeric attribute into an ordinal attribute.

Example :- Temperatures Ranges.

Low (1-15°C) → Medium (16-30°C) → High (31-45°C)

★ Numeric Attributes :-

Quantitive attributes that are measurable. arthemitic opera

★ Internal Scat Att

Key-characteristics :-

→ No obsolute "zero"

Measured in equal size Scale

★ Interval Scale-operation

→ operations that are meaningful diff b/w values

Comparison of order

Two Types :-

Interval Scale Attributes

Ratio Scale Attributes

Examples :-

⇒ Date — No zero "date" (X) and Ratio are not meaningful

Temperature (°C, °F) — Exp 20°C is not "twice as hi as 10°C"

Can +ve, 0 or -ve (Years)

★ Ratio Scale Attributes

→ has meaning obsolute zero points

→ Support all mathematic operation.

→ Ratios b/w values are meaningful and interpretable.

Examples :-

→ Length : 10 meters is Twice as log as 5m

→ Weigth : 80kg is 25% heavier than 64kg

→ Time : 4 hours is half duration of 8h

Lectures No#4+5

Topic:-

Understanding Statistical description In Data Science

we have study different attributes of data in previous class and now this class leeture we have study how "data is desripted" in one Coloumns or values, so we use this Techniques

★ **Data Distribution:-**

- How data is spread across different values.

★ **Data Set Understanding:-**
Methods To gain insight from your data.

★ **Basic Statistical Description:-**
Tools To summerize and interpret data patterns.

★ **Why Statistical Description Matters:**

★ **Data Shape**
Understanding overall distribution pattern of your datsets (Averag Values)

★ **Central Values**
Identify the typical or representative values in your data

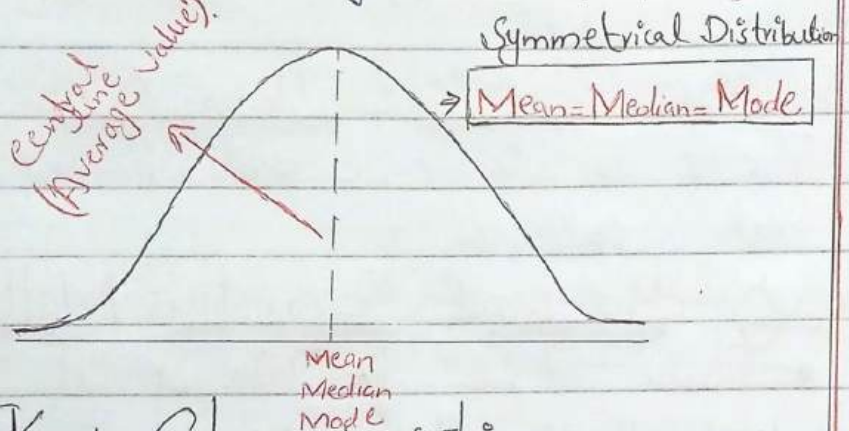
★ **Data Spread**
Measure how wiedly values are distributed from Central Points. (Mean, Median, Mode)

★ **Noise and biase Detection:-**
Spot Random Variation and systematic errors in your dataset

(Types of data Shap)

★ Symmetric data Distribution :-

Definition :- A form of distributed data where Left and Right Parts are Symmetric [Equal]

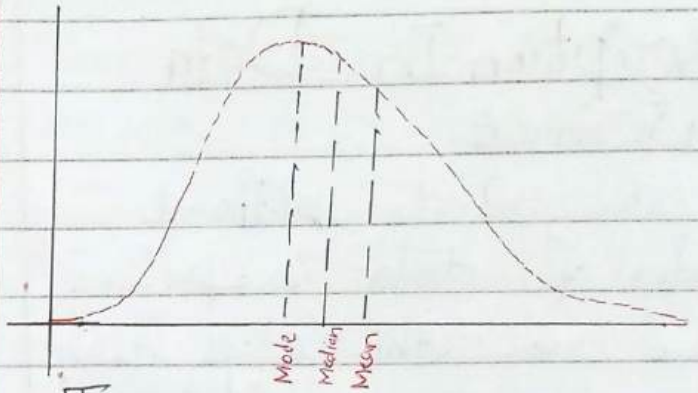


★ Key Characteristic :-
Data is Symmetric around the values.

★ Contrast :-
The alternative form is called Skewed distribution.

★ Skewed distribution

Definition :- A data distribution that is not symmetric in shape.



★ Types :-
Two Types positive & negative Skewed distribution.

Positive VS Negative

Neither type is Symmetric around the central value.

Binary Attributes:-

→ It is a Special types of nominal Attributes only Two values.

Examples:-

→ Gender : Male / Female

→ CoviD : Positive(1) / Negative(0)

⇒ Smoking status : Smoker or non smoker

Sub-Types:-

→ Symmetric Binary Attributes:
Both values are equally important
Example:- Genger (Male / Female)

→ Asymmetric Binary Attributes:-
Example:- Smoker (1 = Smoker, 0 = Non-smoker)
One value is more importan

★ Interchange able words (class, catorey, Lable)

Ordinal Attributes:-

Attributes whose values possess a meaningful order or Ranking between "categories" allowing for comparative relationships.

IMPORANT Characteristics

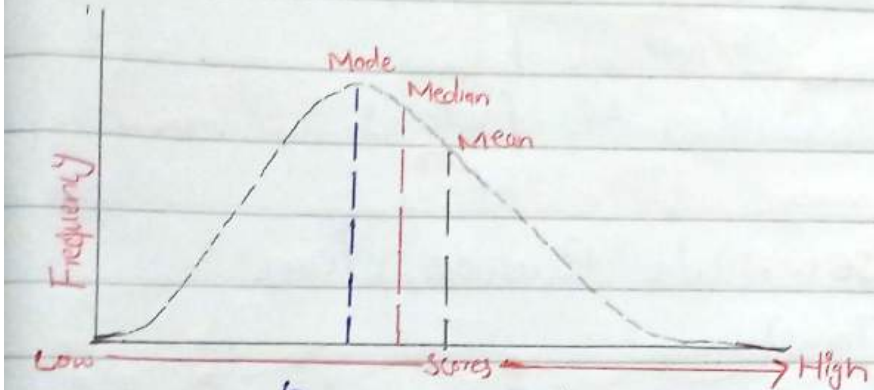
while we can determine if one value is Greater Than, other, the magnitude of differences b/w values cannot be precisely quantified.

Examples:- Academic Grade A+ (ellenct → A (very Good). B (Good). ——
Rands : Assistant Professor, Associate → Professor, full Professor.

Represent:- Code, Seimple not numerical values Two features most frequent values is (modes) :: Between values (Median), Extract from numerical (Not perform arithemtic operation)

Positive Skew (+)

- Data Spread at start, curve goes down slowly. Tail on Right side.

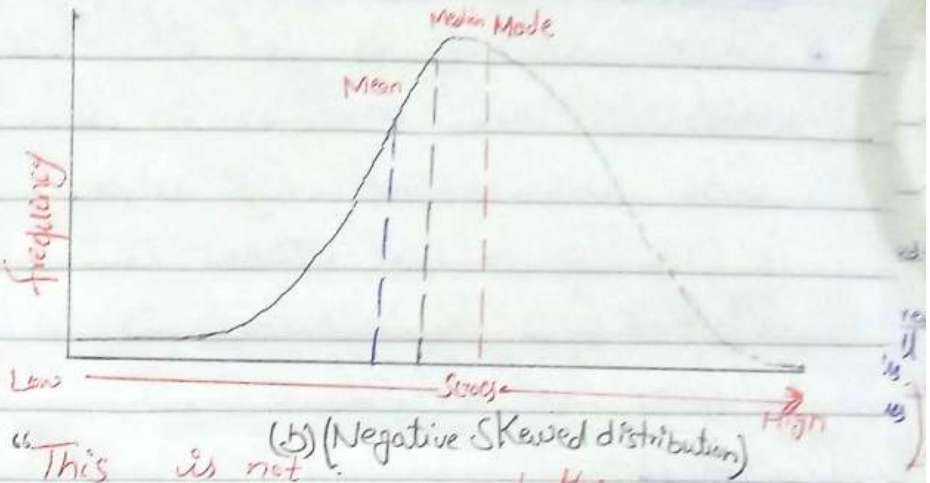


(a) (Positive Skewed distribution)

Negative Skewed :- (−)

- Curve gradually increase, than abruptly closes Tail on Left Side

value کی start سے Left بڑھتی ہے اور
آہستہ آہستہ کم ہوتی ہے



(b) (Negative Skewed distribution)

"This is not Symmetric central around the values."

★ Basic Statistical distribution Measures.

→ Measure of Central Tendency
→ Measure of disperication
→ Measure of Approximety.

They all give good idea of data distribute.

→ Noise data :- Random value

Examples- (online Quiz)

Take on line Quiz:- 1 your Quiz 1 or person 1 same but marks less due to noise

you complete an Online Quize get full Number but some one distribe, you like play song loudy!

→ Skew/Biase data :-

Means data is leaning in one side (ڈیٹا ایک طرف جھکا ہوا ہے)

particular side pr data zyada ho

Exp:- student weight Rang (5'4 to 6'3")

Add 8 inches in mesurment all

All values are skewed Towards higher number (8). effect all data points

OutLiers in DataSet:-

- Some values in out of Range which is not mention in data like one person Age column is (40 To 70) and others is fallal is (5 To 20) which is outliers in datasets.

Cleans:-

- Remove outliers and errors To ensure your analysis reflects actual patterns not Random Variation.

Measures :-

- The are Three Main Types of Statistical description.

1) Central Tendency (2) Dispersion (3) Similarity

Central Tendency — Measures of Average or middle values (Mean, Mode, Median, Trimmed Mean)

Similarity — Measure of Relationships b/w features. (columns) → numeric attributes

Use:- Data Reduce, clean

Dispersion — Measures of data Spread and variation

Transforms:-

Measure of Central Tendency:-

A Measure of central Tendency represent a Single value that describes the centre of dataset.

Variable, features, attributes which is divided into Two Parts (50% above, 50% down values)

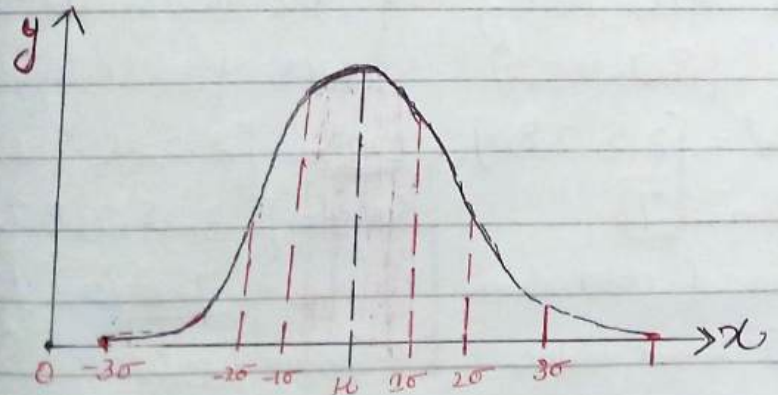
★ Data Divison

It effectively divides the datset into Two equal (halves)

★ Typical value

Indicates the most common or average value in the distribution.

Exp :- In a class where mast student are similar height that common height Represent the central Tendancy.



Means (Simple Average) (Arithmetic Mean).

Sum of all values. divid by observation.

Formula :-

$$\rightarrow \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$\sum x$ = Sum of all values
N = Num of values

Key Points :-

- Simple To Calculate
- highly Sensitive To extreme values
- Used when data has no extreme value.

Example :- Salary of Employees.

Pos.	Cons (کمزوری)
→ Esay To calculate	→ very sensitive To extrem values. (اگر ویلیوز کی شدت زیادہ ہو تو)
→ Cost Effective : Efficient computation power in every data	(اگر ویلیوز کی شدت کم ہو تو)
→ Comprehensive all calculation in every data pts	→ When we used in data no extreme values is Present.

Trimmed Mean

A Refined Average of Means.

is a modified mean where the highest and lowest values are Removed befor Calculation.

Remove Extreme Values. (4 person Salary) Average and "5 hightest Salary is Removed.

How Much You Trimmed Values

As a Rule, do not Trim more than 2% of the Total data points.

Median: (Middle Value) درمیانی ویلیو

The median is the value that lies exactly in the Middle of Sorted dataset. (Ascending order)

Usually it is present in data. if you known in variable you can Able do 50% Lower and 50% higher value differenate b/w them.

Calculate:-

Case 1:-

when odd numbers

Select the middle value after Sorting

Exp:- [8, 5, 9, 3, 7]

Sorted = [3, 5, 7, 8, 9]

Median = 7

Numbers of observation is '5' Single value

Case 2:-

When Even Numbers

Take the Average of the Two middles values after Sorting.

Exp:- [8, 5, 11, 9, 3, 7]

Sorted = [3, 5, 7, 8, 9, 11]

$= \frac{7+8}{2} = \frac{15}{2}$

Median = 7.5

Advantage
- not sensitive To extreme values.
- Use in data pre-processing and cleaning

Disadvantage:-
- Costly To Calculate for lag data
- Limited data Usage. only middle values.

(Advantage)
- It Apply all Types of data
- not affected by extreme values.

(disadvantage)
- not exist in some dataset.
- Less useful for continuous data
- May not represent centre Tendency well

Mode (Most Frequent values)

- The mode is the value that appears most frequently in a datasets. It can be used with numerical and categorical values.

Example:-

dataset = [1,2,2,5,7,8,8,9,10]

Mode = [2,8]

- Mulliples values Repreats and May be no mode in datasets.
  - Multi Model data.

Midrange: (Compulation Purpose)

- The midRange is calculated as.

$$\frac{\text{Minimum value} + \text{Maximum}}{2}$$

Provides a very quick and Rough estimate of the centre. Not Suitable for detailed analysis but for initial approximation

Data Shape:-[distribution]

Symmetric	Positive Skew	Negative Skew
Mean = Median = Mode	Mode < Median < Mean	Mean < Median < Mode

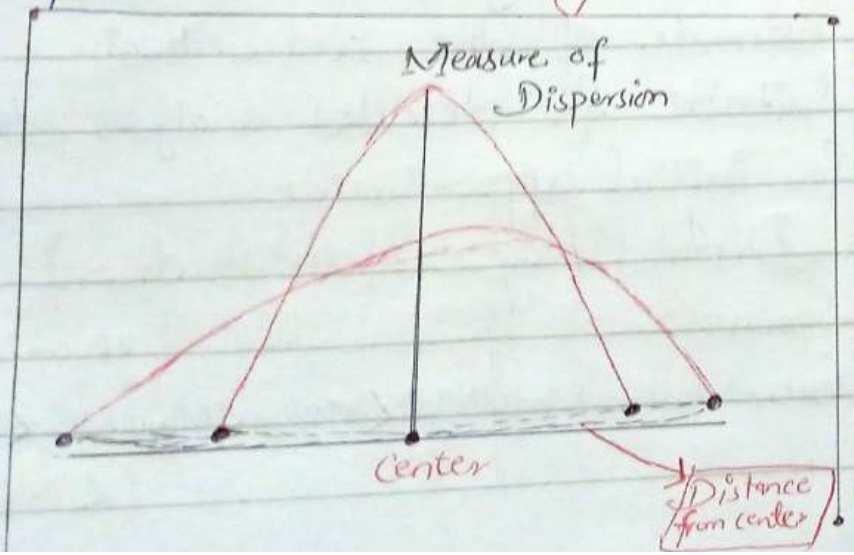
Lecture No#6

Topic:-

Measure of Dipersion (12/11/25)

Define:-

- It refers To how spread out the data is. It shows how far values lie from the central value (Mean or Median)
- Understanding dispersion helps us interpret the variability in data



→ Why is Dispersion Important

→ Beyond Central Tendency
It alone does not show data variability

→ Different Spreads
Two datasets with the same mean can have very different Spreads

→ Added Context
Dipersion adds context to the central value.

★ Choosing the Measure of Dipersion:

→ Median based Measures
If we use Median we use quantiles and interquartile Range

→ Mean based Measure
if we use Mean, we use standard deviation and variance:

→ Selection Criteria:-
Choose depend upon the Measure of central Tendency used

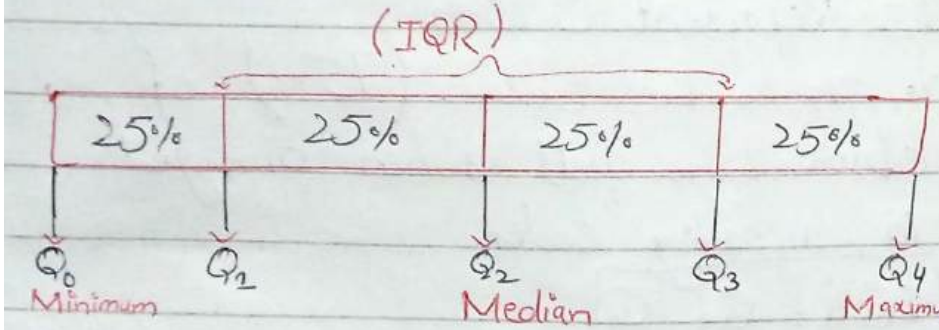
Assuming :-

*(Dipersion from the Median) :-

- When the median is the central measure.

→ **Quantiles**
We use quantiles (e.g- quantiles, percentiles)

→ **Interquartiles Range**
We focus on interquantiles Rang (IQR)



*Understanding Quantiles :-

(1) **Definition:-**
are the values that divide a dataset into equal-sized groups or portion

(2) **Divison :-**
When we divide data into 'q' equal parts, we get (q-1) quantile values as boundaries b/w these parts

(3) **Median**
When $q=2$ we get 1 quantile value → This is the Median, dividing data into 2 equal halves

(4) **Quartiles.**
When $q=4$, we get 3 quantiles value → These are Quartiles dividing data into 4 equal quartes.

★ Median as a Quantile :-

→ 2-Quantile:
The median is 2Q especially, representing the middle value that divides a Sorted dataset

→ Equal Halves
it perfectly splits the dataset into 2 equal halves with 50% of values below and 50% above.

→ Data Spread :-
⊙ The median provides crucial insight into how values are distributed on either side of the central point revealing the dataset's balance.

1 1 3 5 7 8 10 11 11 15

25% 25% 25% 25%

1st Quartile — median = 2nd Quartile — 3rd Quartile

★ Quartiles Explained :-

→ Q1 first (Quartile)
Q1 Represent the 25th % the value below which 25% of observation fall when arranged in ascending order

→ Q3: (Third Quartile)
Q3:- Represent 0 75th % the value below (75%) of observations fall when arranged in ascending order.

→ IQR = $Q_3 - Q_1$

→ Q2 (Median)
Q2:- Represent the ~~50th~~ % (Me the Middle value dividing the dataset into Tw equal halves.

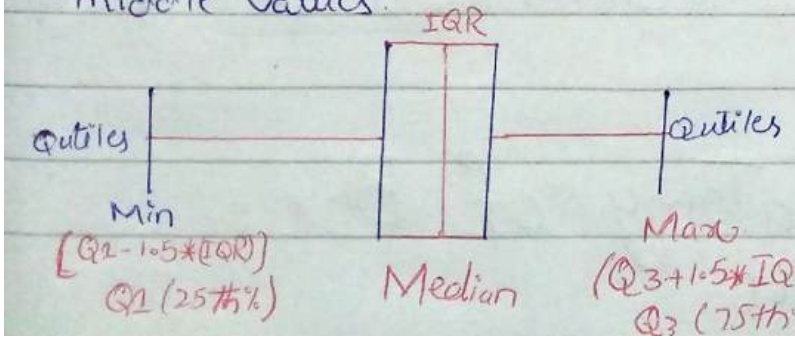
(IQR):-

- Measure data dispersion by fousing on the "middle values, eliminating the influence of outliers."

i) Definition & formula:
IQR represent the middle 50% of data Points. $IQR = Q_3 - Q_1$
Q3 is the 75th% and Q1 is 25th%

ii) Data Concentration
IQR measure the spead of the Middle 50% of observation for extrem values. (Small IQR)

iii) Interpreting IQR values: Tightly clustered middle values Large IQR widely spread middle values.



*5-Number Summary:-

- Give a complete picture of distribution and spread.

→ Minimum
The Smallest value in the dataset

→ Q1 (First Quartile)
25th% of the data

→ Median(
50th % percentile
Middle valu

→ Q3 (3rd Quartile)
75th% percentile of the data

→ Maximum
The largest value in the dataset.

Note:-
if you know the any Numeric feature we have use it To find.

- Central Tendancy of data
- data Spread throug IQR
- Overall distribution of Shape.

Samply used IQR quantiles.

★ Understanding data Shape:-

◎ The data is skewed if more spread on one side of Median

i Positively Skewed
$Q_3 - Q_2 > Q_2 - Q_1$

ii Symmetric
Data is symmetric
$Q_3 - Q_2 \approx Q_2 - Q_1$

iii Negative Skewed
$Q_2 - Q_1 > Q_3 - Q_2$

★ Outliers :- What Are They?

◎ Are the extreme values outside the normal Range. They can distort analysis and give misleading insights.

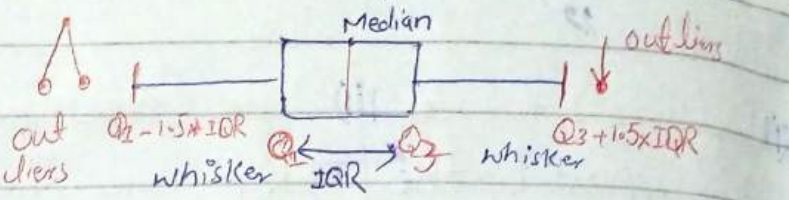
★ Identifying Outliers:-

→ Calculate IQR
$IQR = Q_3 - Q_1$

→ Lower Bound
Lower bound
$Q_1 - 1.5 * IQR$

→ Upper bound
◎ Uper bound
$Q_3 + 1.5 * IQR$

→ Identify
◎ Values beyond these are outliers



★ Box & Whisker Plot for Visualization.

★ Complete Visualization
visual Representation of [Min, Q_1, Median, Q_3 Max] Also Showers outliers

★ Comparison Tool
Multiple distribution for Comparing.

★ Anatomy of Box :- Plot

→ The Box
Box show IQR (Q_1 to Q_3)

→ The whiskers
Extent To Min and Max (non-outliers) values

→ Median Line
Line inside is the 'Median' (Q_2)

→ Outliers
Dots show outliers.

★ Variance:
It measure the Average Squared deviation from Mean.

→ $\sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2$

★ if Measure is Means:-
ⓐ Using this Standard Deviation
(i) Square Root of Variance.
is the Square root of variance

→ $\sigma = \sqrt{\sigma^2}$ اس کی effect (+ve) اور (-ve) observation پر ہوتی ہے۔

(ii) Deviation Measure
indicates how much data deviates from the mean

→ The Box
Box show IQR (Q1 to Q3)

→ The whiskers
Extent To Min and Max (non-outliers) values

→ Median Line
Line inside is the 'Median' (Q_2)

→ Outliers
Dots show outliers.

★ Variance:
It measure the Average Squared deviation from Mean.

$$\Rightarrow \sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2$$

★ If Measure is Means:-
◉ Using this Standard Deviation
(i) Square Root of Variance.
is the Square root of variance

$$\rightarrow \sigma = \sqrt{\sigma^2}$$

(ii) Deviation Measure
indicates how much data deviates from the mean

Lectures No#7+8

Topic:- Proximity & Similarity

★ Introduction:-
◉ In data Science, understanding the Relationship b/w different features or Attributes/columns of a dataset is crucial for analysis and model development.

★ Proximity and Similarity:-

- Measures are Techniques used to quantify how close or similar Two or More Attributes are.

★ Importance of Similarity in DS

- **Improved data understanding**: its Similarity improves data understanding.
- **Preprocessing Guidance**: This information guides data preprocessing & modeling
- **Numeric Attribute focus**: It is especially imp for Numeric attributes.
- **Feature Relationship**: It identifies which features are related or independent.
- **Transformation Impact**: Directly impacts the effectiveness of Transformation Techniques.

★ Representing Attributes as Vectors

- **Vector Representation**: we represent each attributes in vectors
- **Mathematical foundation**: The representation conver the rich Theory of Linear algebra
- **Similarity calculation**: Using vectors, we can easily calculate Similarity
- **Structured Approach**: Vectors provide a Structured way to handle Numeric data
- **Operationaly flexibility**: It allows us To apply various Mathematical operation b/w attribues

★ Different Proximity Measures:-

- **Dot Product**: we will begin with the dot product (·)
- **Convariance**: Next, we discuss it $(-\infty, +\infty)$
- **Cosine Similarity**: we more To the cosine Similarity
- **Correlation**:- finally we introduce Correlation for evaluating strength and direction. Rang (-1 to 1)

★ Similarity Ranges:-

(1): vector are perfectly aligned
(0): vector are orthogonal (no similarity)
(-1): vectors are perfectly oppositive.

★ (•) Dot Product :-

→ It Takes Two vector (attributes) as input and give output as number. If [0] numbers There is minimum difference Similarity. If [num > 0] then more chance in Similarity.

"دو (2) Vectors کا جوائنٹ push کتنا زیادہ ہے
جتنا زیادہ ہوگا اتنا ہی زیادہ دونوں ویکٹر ایک سمت
میں push کر رہے ہیں"

★ Pros & Cons :-

Pros	Cons
→ Simple Calculation: Simpe To calculate and computationaly efficient	→ Basic Measure: It provide a basic Measure of the Relationship b/w vectors
→ Magnitude (strength) dependency: Its calculation depend on the vector magnitude	→ Unit :- changing unit of Measure can Result in drastically diff(.) Values.

⊛ Cosine Similarity :-

→ Direction focus: focuses Solely on the direction of vectors

→ Magnitude direction: Reslove the magnitude issue of dot product

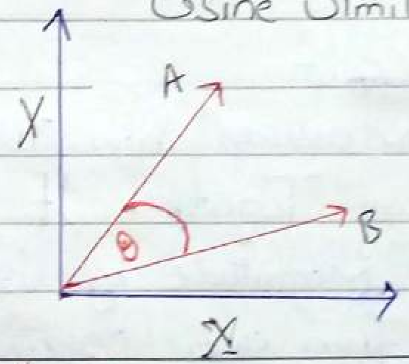
→ Perfect alignment

→ Vector Normalization: Normalize vector To unit vector

→ Bound Range: Rage b/w (-1, +1)

(+1) implies perfect alignment and (-1) implies opposite

Cosine Similarity



→ $Sim(A, B) = Cos(\theta) = \frac{A \cdot B}{|A||B|}$

Benefits :-

- its Magnitude b/w (-1 to +1).
- if Cosine Similarity (-1) The both attributes is different.
- if Cosine Similarity (+1) Same direction attributes of both sides. (Trends Same) hy.
- if Cosine Similarity is '0' both is independant (no Relation).

Covariance :-

- Covariance Measures how Tow variable Vary Togather. It Consider th Magnitude by (Subtracting the mean from each observation vector)
- (+ve) Convariance indicates that both attributes Trends To Move Togather.
- (-ve) indicates that They move in oppositive direction

Formula :-

- Convariance formula b/w Random variable X, and Y.

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

where

- n = observation Num
- x_i = i-th value of variable x
- y_i = i-th value of variable Y
- $\bar{x}$ = Mean of X vector
- $\bar{y}$ = Mean of Y vector

Range = $(-\infty, +\infty)$.

⊛ Cons:-

- Direction of variable provided but not strength is not provided.

Correlation:- (Mostly used for Data Reduction)

- Perfect Relationship
  - (±1) indicates perfect relationship
- Standardization
  - Derived by dividing Convariance by Standard deviation
- Bound Range.
  - value bounded b/w (-1 and +1) (-1 opposite, +1 same)
- Relationship Measure
  - Quantifies both Strength and direction

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma X \; \sigma Y}$$

"X by variables and ÷ by Standard Variation and atthragh is variance"

⊛ Summary :-

The Summary of Proximaty.

Measures	Characteristics
Dot Product	Simple operation, Magnitude dependent.
Cosine Similarity	Normalize vectors, focus
Covariance	Incorporates magnitude and direction but unbound
Correlation	Standardized version of Convariance b/w (-1, +1)

Topic:- Data Quality & Pre-Processing

Data Preparation Components:

(1) Exploratory Data Analysics
(2) Data Preprocessing

- Data کی کوالٹی کیا ہے؟
- ٹیزی کب بہتری ہے اور سے اچھا کر سکتے ہیں؟ کیوں ضروری ہے؟

Importance of Data Quality in Preprocessing.

Essential Step befor modeling.

- Good Models
  - Successful Machine learning
- Essential foundation
  - before modeling begins.

- Good data
  - Quality preprocessing
  - Good data → Good Models
- Poor data → Poor Prediction
  - ("Garbage In, Garbag out")

Real-World Data challenges

- Missing values.
  - Incomplete Reord requiring handing
- Inconsistent formats
  - Varying Standard ceross reords.
- Duplicates entries
  - Redundant information needing removal.
- Irrelevant features
  - Unnessary coloumns

★ Kaggle:-

- Using dataset and get it websites all issues in actual data.

★ Data Quality:-?

- Accuracy of values
  - Ensuring data Correctly represents Reality
- Minimal Missing values
  - Complete inform for analysis
- Consistency in Attributes.
  - Uniform formats and standards ceross dataset
- not duplication:-
- Timelines and Relevance: Update inform from current needs

★ Preprocessing of Data:-

- To improve the quality of data is called preprocessing of data which give insight of business help.

Clean	Integrate	Reduce	Transform
→ Remove errors & inconsistencies	→ Combine Multiple Resources	→ Reduce volume while Maintain quality	→ Convert To Suitable format for

Data

★ Techniques for Data Reduction

Several Methods use.

→ Feature aggregation

- Combining Multiple attributes into meaningful Summary.

→ Feature Selection:

- Identifying and retaining only the most Relant based variables.

→ (PCA: Principle Component Analysis):-

- Creating new uncorrelated variables that capture Maxi variance in dataset.

→ Dimensionality Reduction :-

- Using T-SNE or UMAP To visualize and compress high-dimension into lower dimesion.

★ Rows Sampling:-

There is no duplication in our datasets. (10 Millions) Rows.

→ Original dataset → Sampling Process
→ Reduce dataset.

★ Data Transformation :-

→ Scaling features To Similar ranges.
→ Making data uniform for algorithms
→ Prevent bias b/w larger Scale feature Scales b/w (0, 1)
→ Helps in faster and better Convergence.

Befor Transform	After Transform
→ features on diff Scales	→ Uniform features Scales
→ Algorithms bias Towards larg Scale	→ Equal important in Algorithms