⑥ Data dis...

① What is data distribution ?

→ A data distribution shows how values are spread in a data set.

→ It tells you where most value lie, how spread out they are, and whether extremes exist.

→ Population Vs Sample :-

① Population :- Entire group you want information about.

Ex :- ① All customers of amazon
② All Students in india
③ All transaction in a year.

→ usually too large, expensive or impossible to measure.

② Sample :- A subset of population, used to estimate population behavior.

Ex :- 10,000 Amazon users.
500 Students from different states.
Last 1 month of transaction.

→ In ML & DS we almost always work with samples.

② Mea...

Ex :- Sala...

→ Mean ... with

Ex :-

→

• Mea...
• In

→ In

→ use...
→ Av...

dist...

(2) Mean :- Mean = Sum of Values / Num of Values

Ex:- Salaries = 20, 22, 25, 30

= (20+22+25+30) / 4 = 24.25

→ Mean becomes completely useless when encountered with outliers.

Ex:- 20, 22, 25, 30, 5,00,000

→ completely useless.

- Mean is sensitive to outliers.
- In skewed data mean lies towards the tail.

→ In M.L Pipelines :-

→ use mean for clean, symetric data.

→ Avoid it for income, sales, prices without checking distribution.

③ Variance :- Mean tells center.
Variance tells spread.

→ Variance is a statistical measure that quantifies the spread or dispersion of a set of data points around their mean (avg).

→ To find the variance we are going to first find the mean of data.

85, 90, 95, 100, 105

→ mean = $(85+90+95+100+105)/5 = 95$

→ Now after finding the mean we have to compute the difference from each point and mean and squaring them.

85, 90, 95, 100, 105

mean = 95

$(85-95)^2 + (90-95)^2 + (95-95)^2 + (100-95)^2 +$

$(105-95)^2 = 250.$

→ Now just divide the computed difference from the numbers of points in your data.

$250/5 = 50.$

→ Now afte
the differen
squaring th
→ formula

Pe

→ formula

→ Now after finding the mean we have to compute the difference from each point and mean and squaring them.

→ formula for population → $\left[\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right]$

Population	Sample	Name
N	n	Total element
μ	$\bar{X}$	mean
σ^2	s^2	Variance
σ	s	STD.

→ formula for Sample → $\left[s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$

→ Standard deviation :-

Now to find the Standard deviation you just have to take the Square root of the Variance.

85, 90, 95, 100, 105

Variance = 50

STD Deviation = $\sqrt{50}$ = 7.071

→ calculating the distance

→ formula for Population :- $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

→ formula for Sample :- $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$