





DAV-3

# HYPOTHESIS TESTING



# Lecture 9: Feature Engineering 1

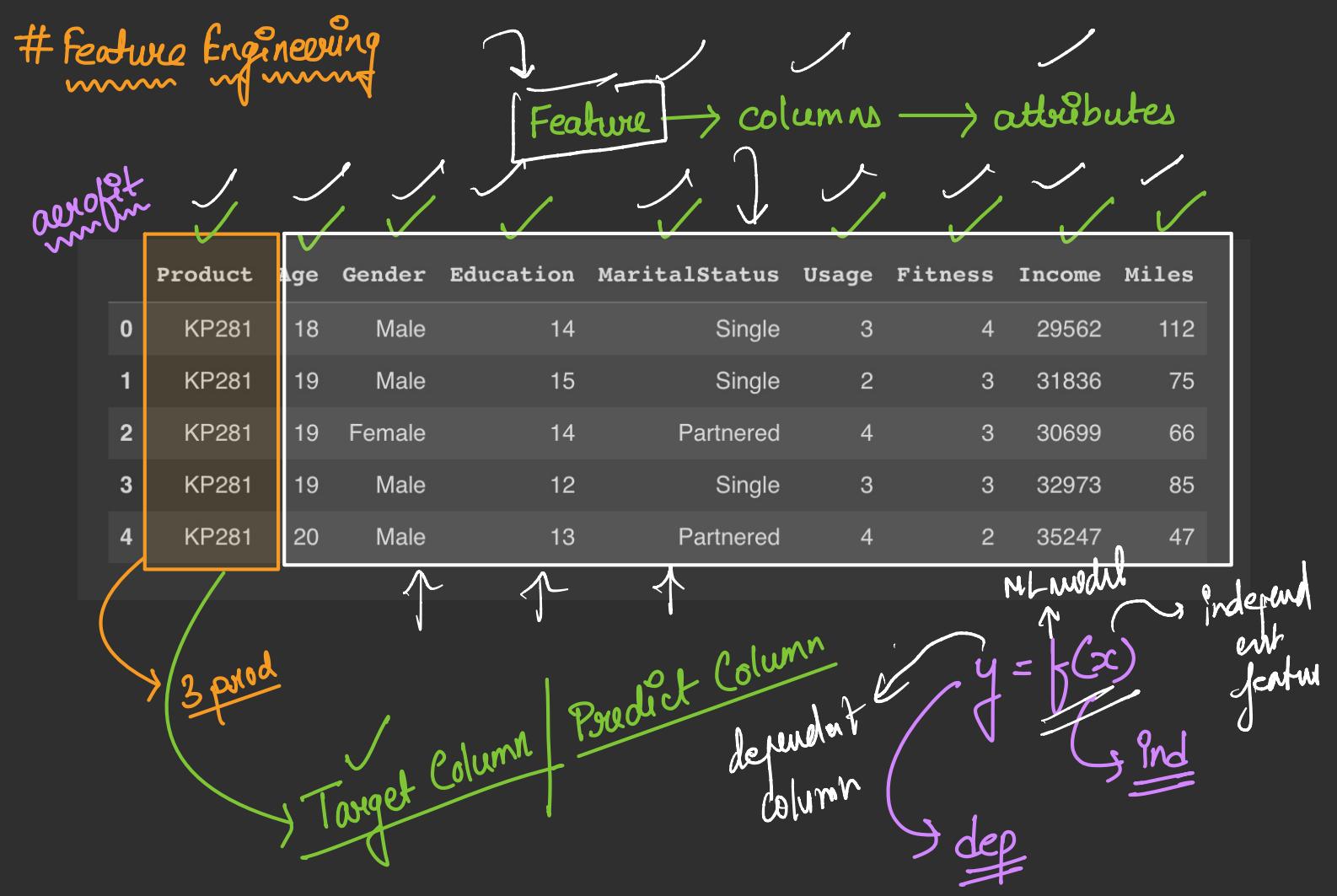
Class starts at 9:05 PM

# Agenda

- 1 Feature Engineering
- 2 Skewness
- 3 Kurtosis
- 4 Univariate Analysis (UVA)

$f \rightarrow f'$   
16/01/2026  $\rightarrow$  year from  
line

Feature what? Engg what? Age  $\leftarrow$  current year - year  
coming up with new set of features  
from your existing data



# Predict the House Prices

Target Col : Sale price of House

Feature : Area, HouseType, Material, Parking . . . .



Area	Balcony	Bedroom
1000 sqft	1	1
2000 sqft	2	2

Feature → predict the Target

## # Fitness Example

Small survey is done → Height & Weight

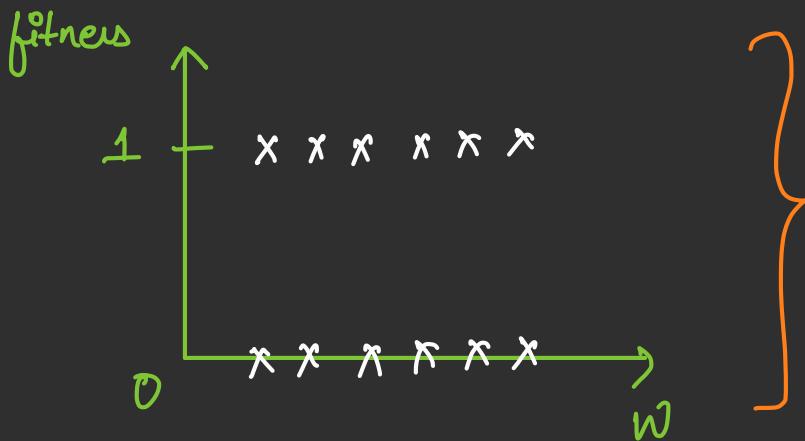
Expert → "SME"

$F_1$	$F_2$	Fit/UnFit
H	W	
—	—	1
—	—	0
—	—	1
—	—	0
—	—	1

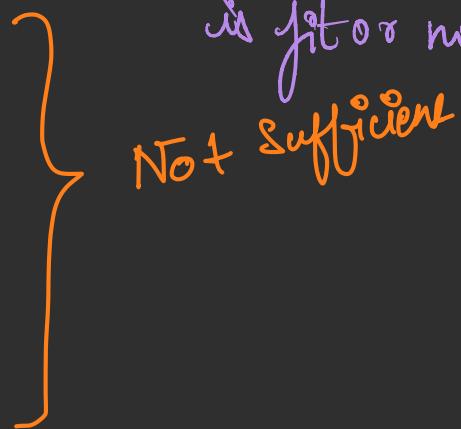
No. of features

Binary Categorical col

labelling of data



The weight feature is  
not enough to tell  
me whether person  
is fit or not



Not Sufficient

BMI  $\rightarrow$  Body Mass Index  $\Rightarrow H/w^2$

derived column

✓ ✓

H	W	BMI	Fit/UnFit
-	-	-	1
-	-	-	0
-	-	-	0
-	-	-	0
-	-	-	1

fitness

1

0

BMI

$f(g - w) \rightarrow 1$

Engg

## # Loan Status Example

Bank → Business → Loan out money

predict  
↓

Q) What are some of the important Feature?

give loan or not

CIBIL Score  
gender

salary

existing loan

arrests

married or not

age

education

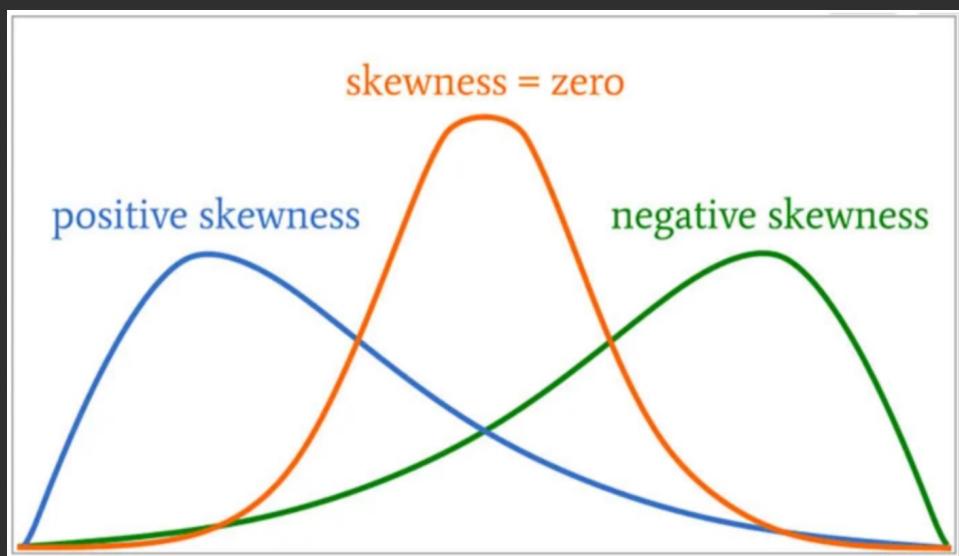
100 → Features  
X

Target

# Skewness

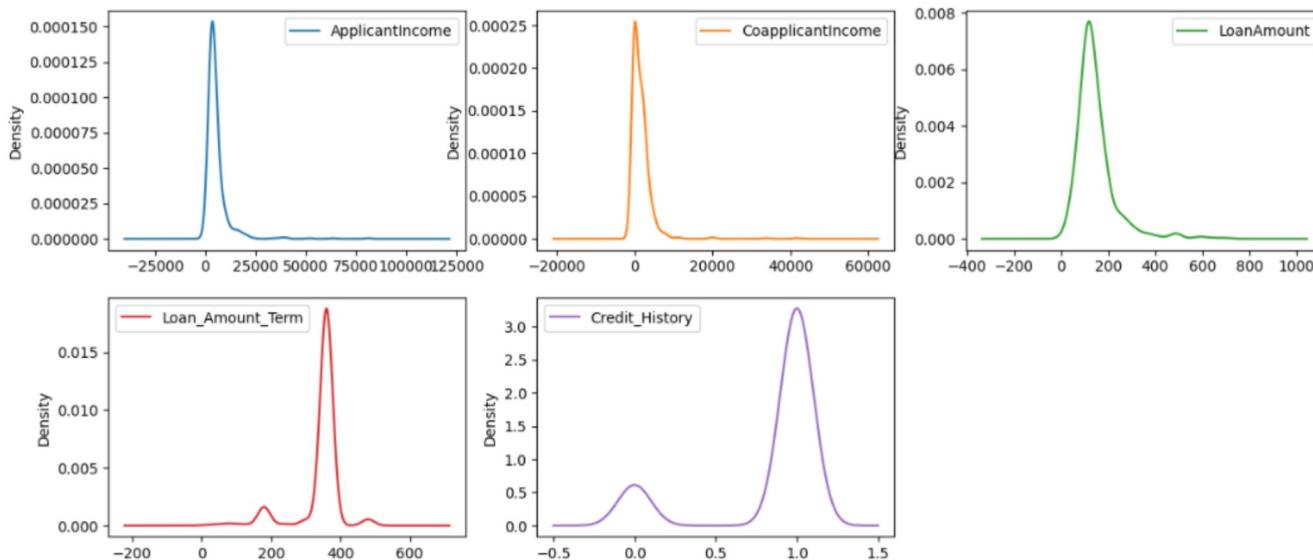
# # Skewness

Skewness measures the asymmetry of a data distribution around its mean, indicating whether data leans towards the left (negative skew) or right (positive skew).



# Skewness → Helps to understand the "shape of the distribution"

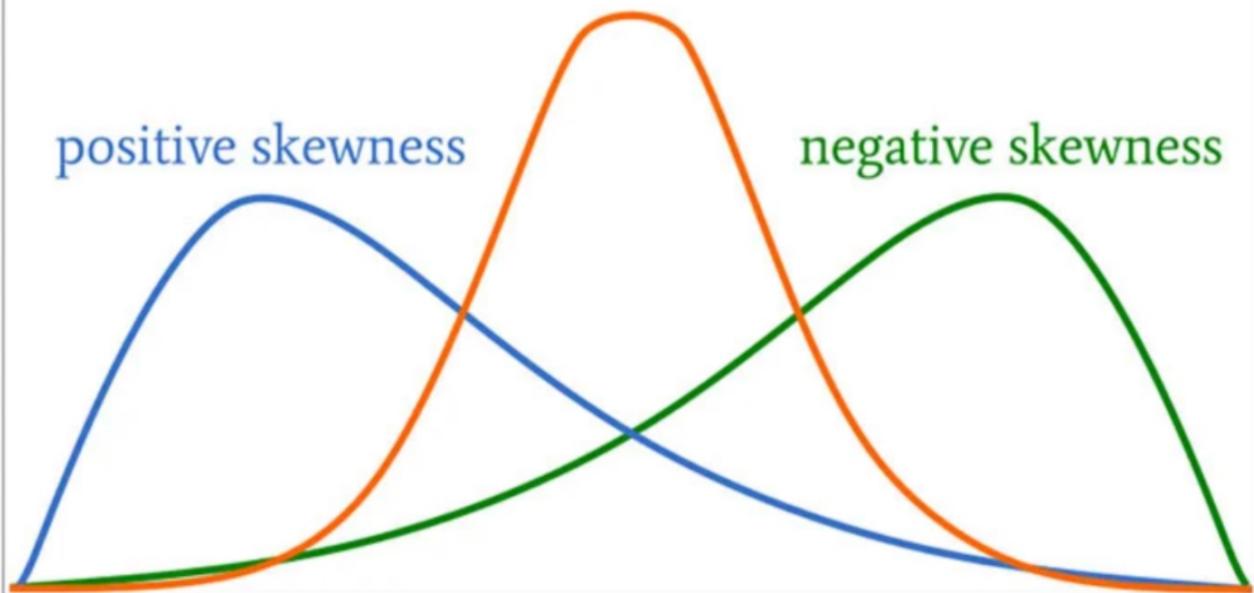
## Output



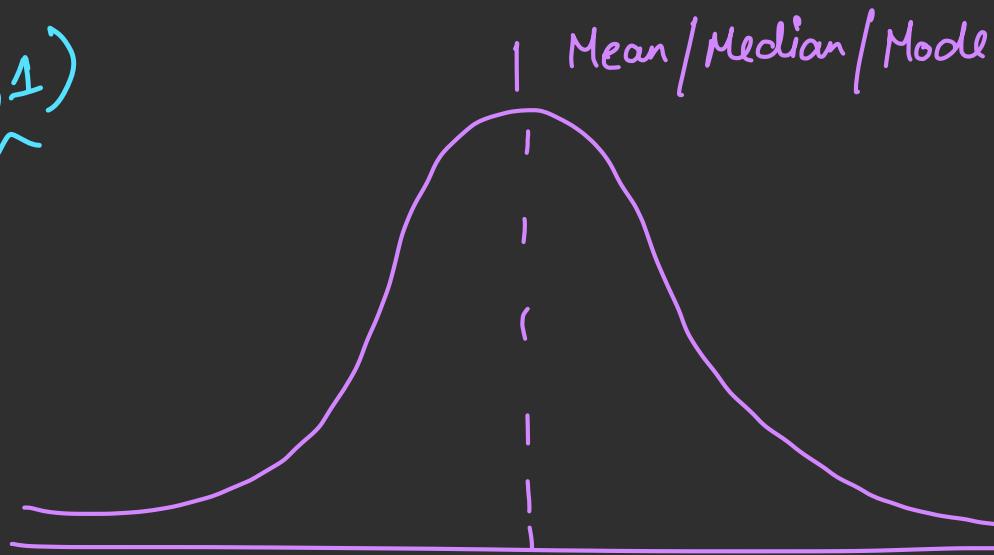
skewness = zero

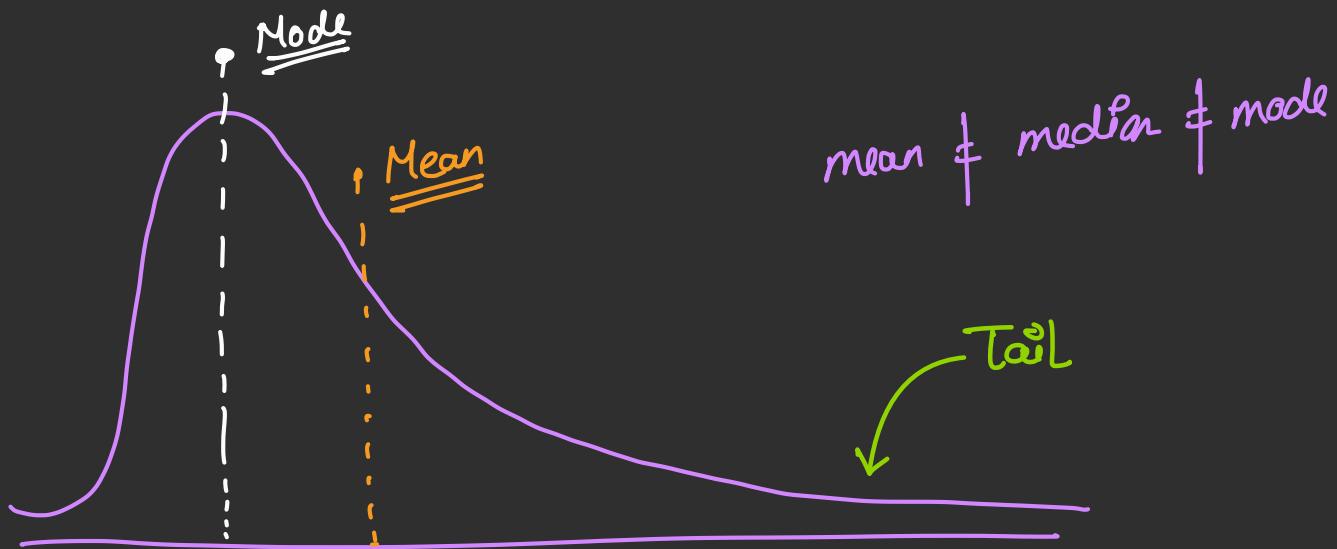
positive skewness

negative skewness



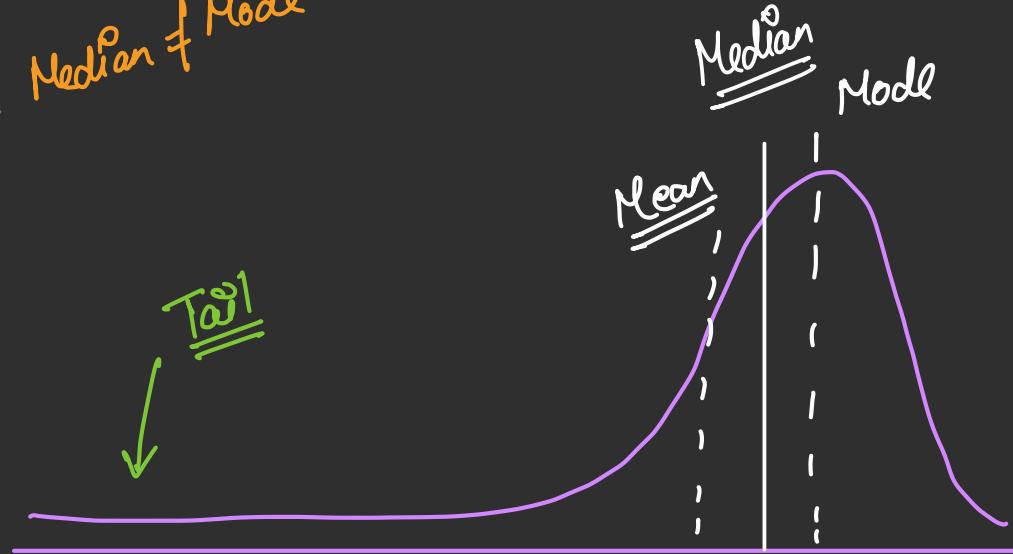
$Z(0,1)$





unsymmetrical  
(positively/right skewed)

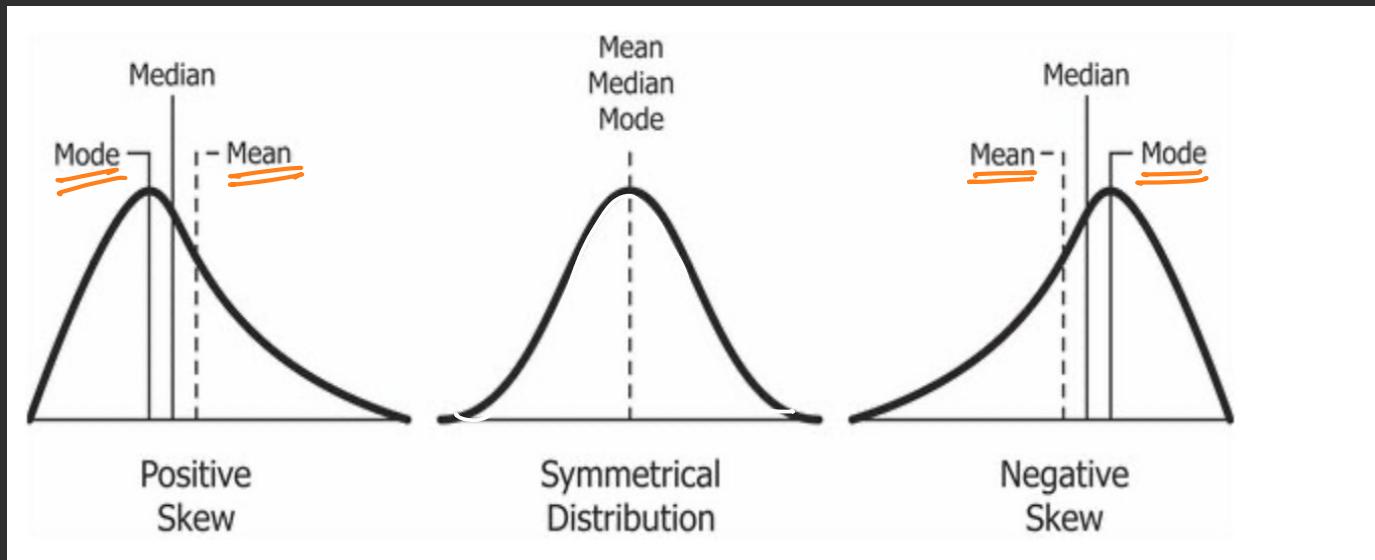
Mean  $\neq$  Median  $\neq$  Mode



(Neg skewed)

(left skewed)

## # Types of Skewness



# # Skewness Formula



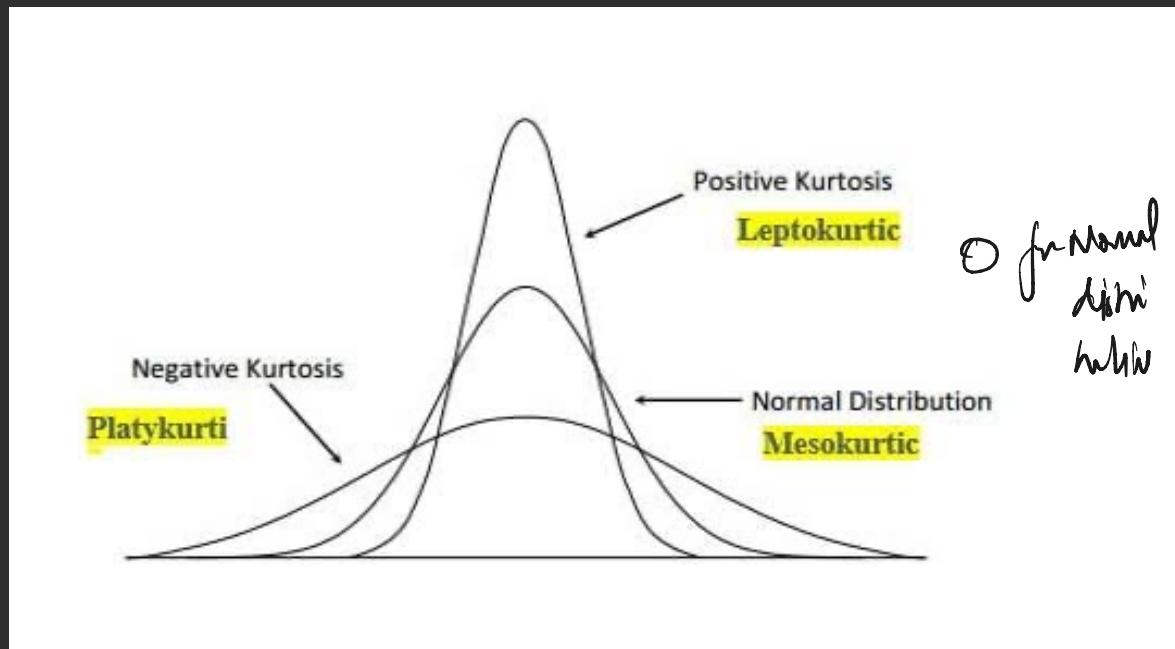
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \text{mean}}{\text{std dev}} \right)^3$$

*z score*

# Kurtosis

# #Kurtosis

Kurtosis tells us how much a data distribution has values far from the average, showing if it has “heavy tails” (many extreme values) or “light tails” (few extreme values) compared to a **normal**, bell-shaped curve.

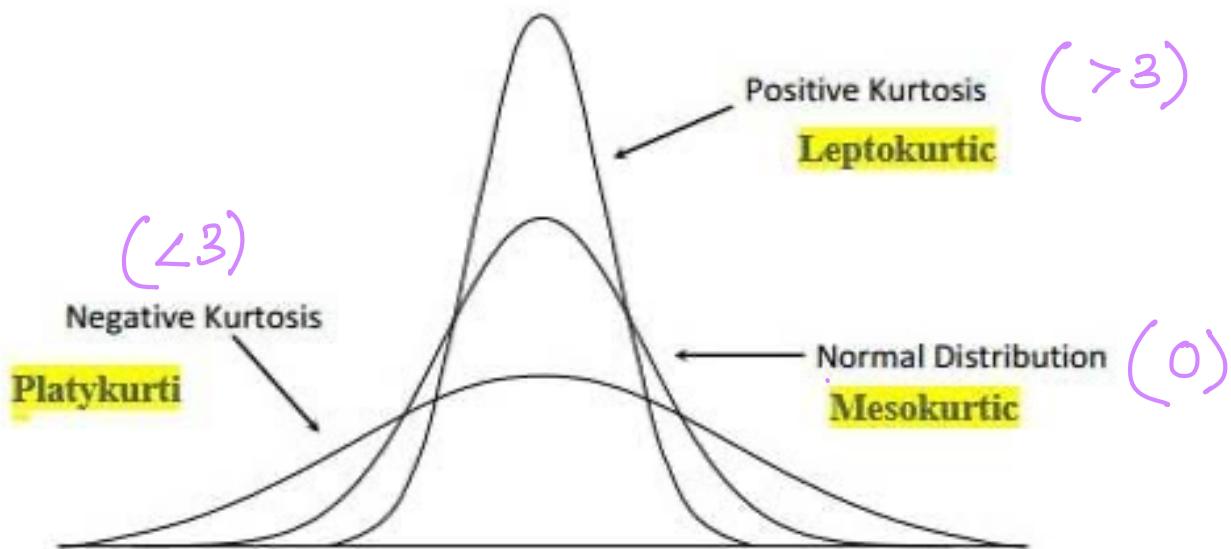


# # Kurtosis Formula

$$\text{Excess Kurtosis} = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \text{mean}}{\text{std}} \right)^4 - 3$$

Kurt

for ND  $\Rightarrow \frac{3}{n}$



## # Relationship b/w Kurtosis & Skewness

### Common points

- ① Both helps us to understand the distribution
- ② They both talk about outliers

### Differences

- ① Skewness → Nature of the dist
- ② Kurtosis → Concentration of data points

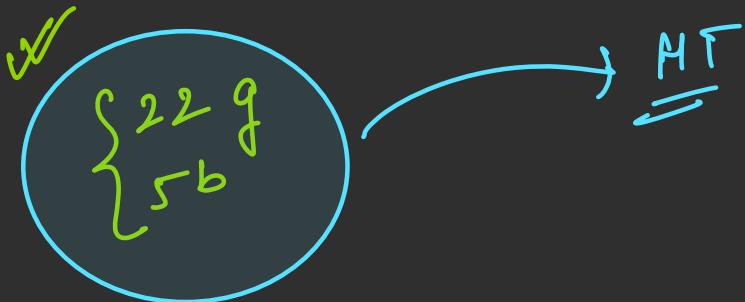
## #Quiz

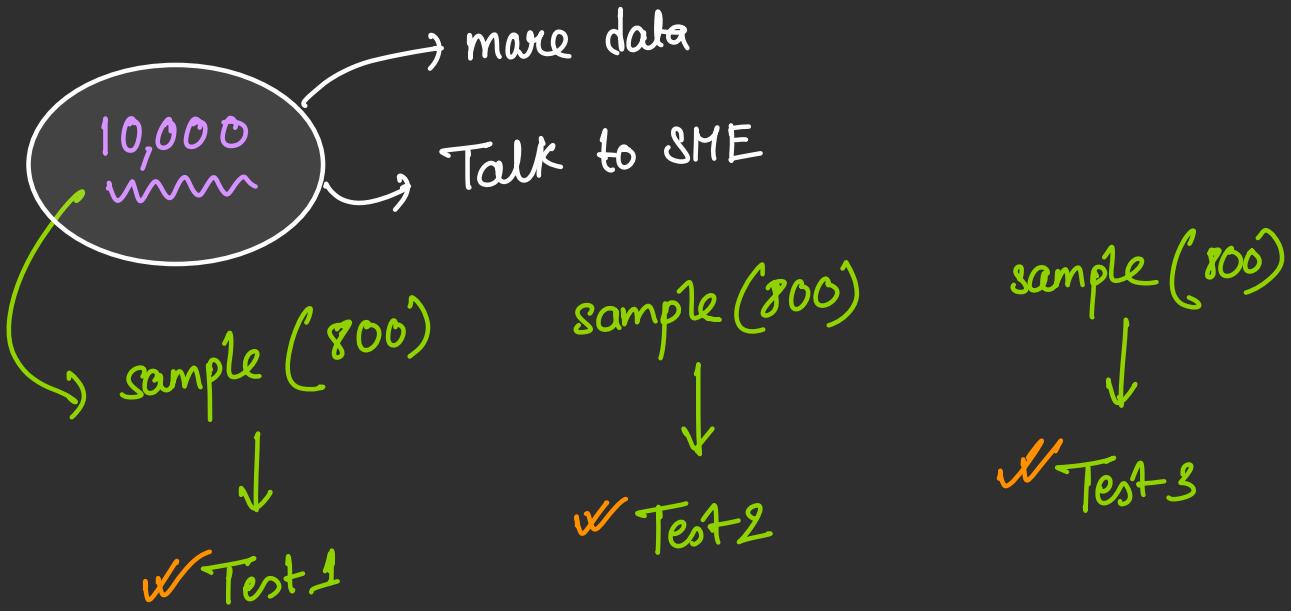
skewness for Normal Dist  $\rightarrow 0$

kurtosis for Normal Dist  $\rightarrow 0$

Both are zero for Normal dist

$g, b, g, b \dots$





## # Action Items

- 1) give me the list of topics, that you are under confident
- 2) get myself to the WA group
- 3) I will provide some reading material by the end of this module.
- 4) Important topics for interview

summary of all Test → shortlisting a Test

- 1) Create a mental map (broad view)
- 2) Practice