



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)

Feature
mm

What?

Engg
mm

What?

?

$f1 \rightarrow f1'$

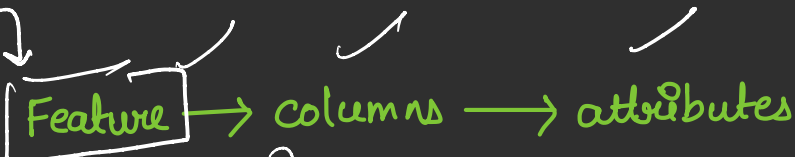
16/01/2026 → year from here

Age years

← current year - year

coming up with new set of features
from your existing data

Feature Engineering



aerofit

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP281	18	Male	14	Single	3	4	29562	112
1	KP281	19	Male	15	Single	2	3	31836	75
2	KP281	19	Female	14	Partnered	4	3	30699	66
3	KP281	19	Male	12	Single	3	3	32973	85
4	KP281	20	Male	13	Partnered	4	2	35247	47

3 prod

Target Column

Predict Column

dependent column

ML model

$$y = f(x)$$

independent features

ind

dep

Predict the House Prices

Target Col: Sale price of House

Feature: Area, House Type, Material, Parking

Feature

Price

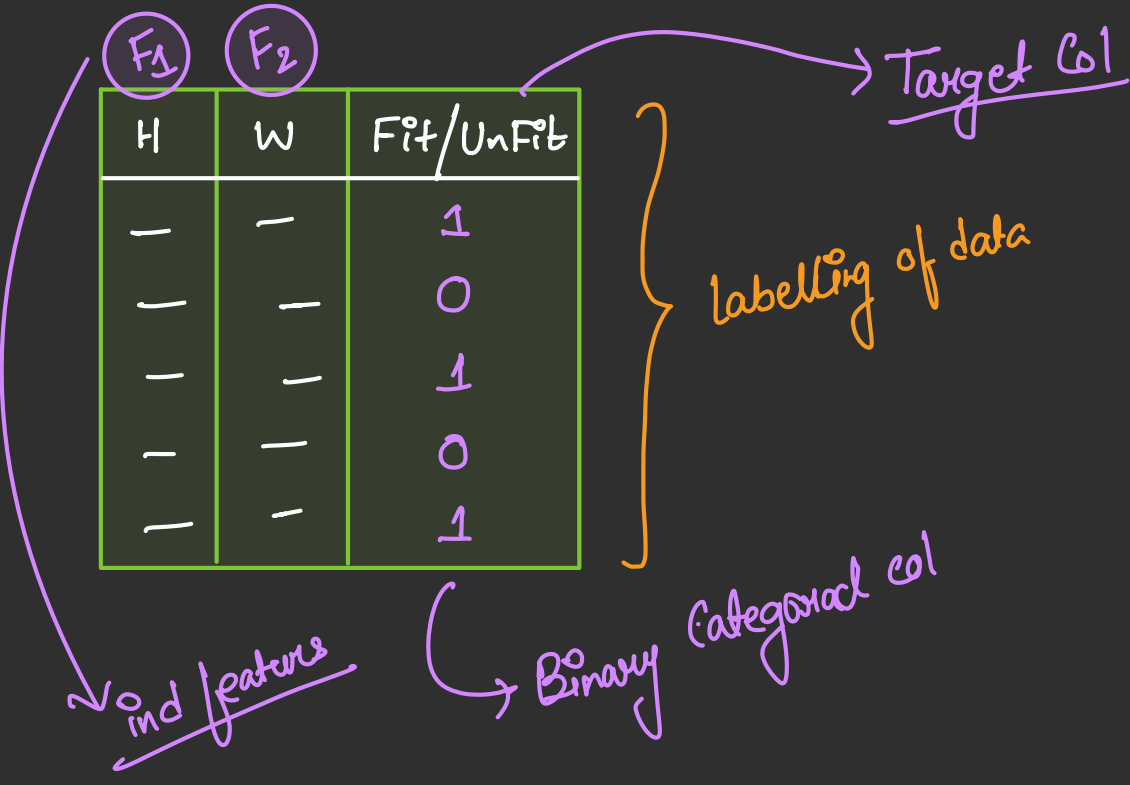
Area	Balcony	Bedroom
✓ 1000 sqft	1	1
✓ 2000 sqft	2	3

Feature → predict the Target

Fitness Example

Expert \rightarrow "SME"

Small survey is done \rightarrow Height & Weight



The diagram illustrates the relationship between the table columns and the handwritten labels. A purple circle labeled F_1 has an arrow pointing to the 'H' column. Another purple circle labeled F_2 has an arrow pointing to the 'W' column. A purple arrow points from the 'Fit/UnFit' column to the label 'Target Col'. A purple arrow points from the 'H' and 'W' columns to the label 'ind features'. A purple arrow points from the 'Fit/UnFit' column to the label 'Binary Categorical col'. An orange bracket on the right side of the table is labeled 'labelling of data'.

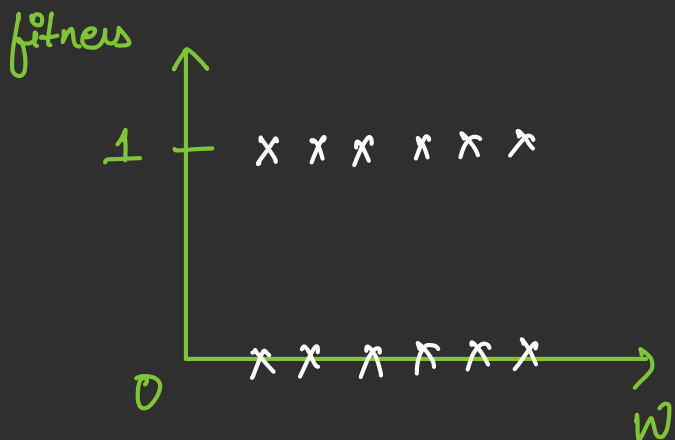
F_1 H	F_2 W	Fit/UnFit
—	—	1
—	—	0
—	—	1
—	—	0
—	—	1

Target Col

labelling of data

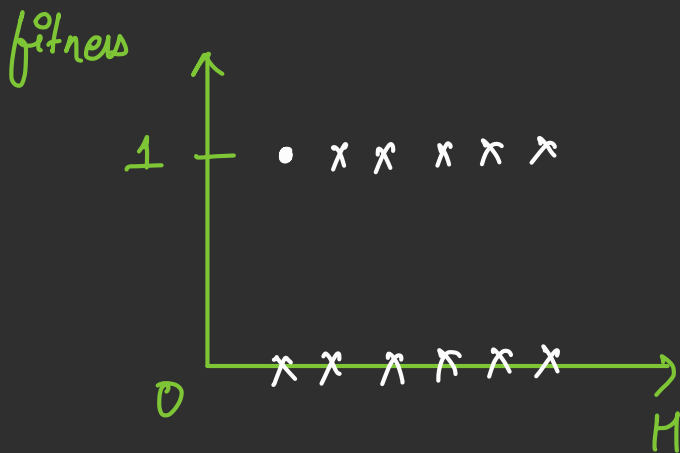
Binary Categorical col

ind features



Not Sufficient

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



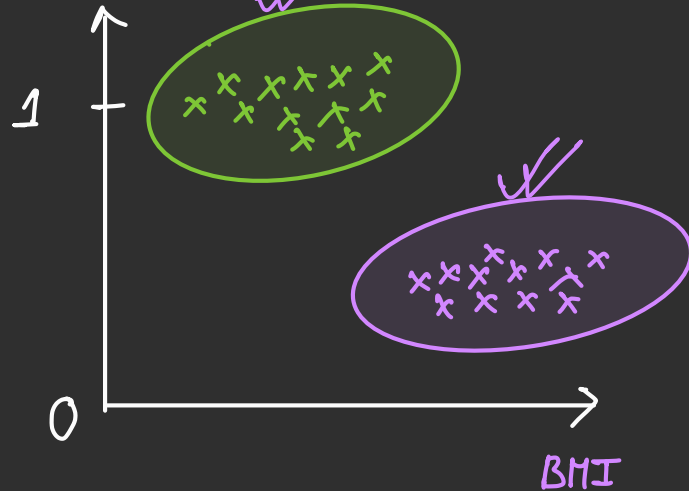
Not Sufficient

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

derived column

H	W	BMI	F _{it} /UnF _{it}
-	-	-	1
-	-	-	0
-	-	-	0
-	-	-	0
-	-	-	1

fitness



19-22 \rightarrow 1

Engg

Loan Status Example

Bank → Business → Loan out money ↓

Q) What are some of the important features?

CIBIL Score

gender

salary

existing loan

assets

married or not

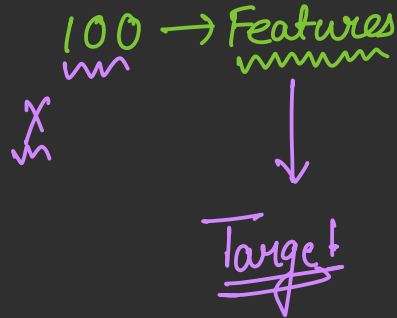
age

education

predict



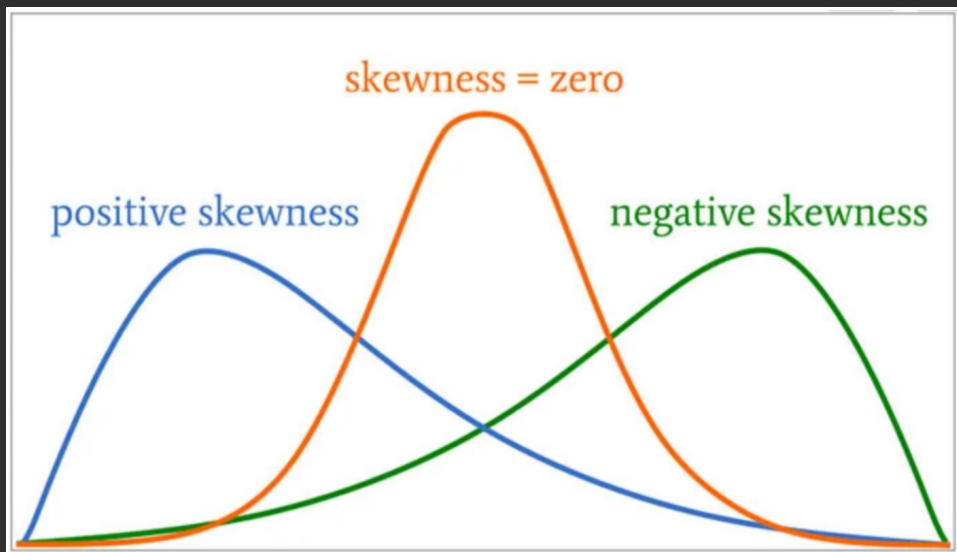
give loan or not



Skewness

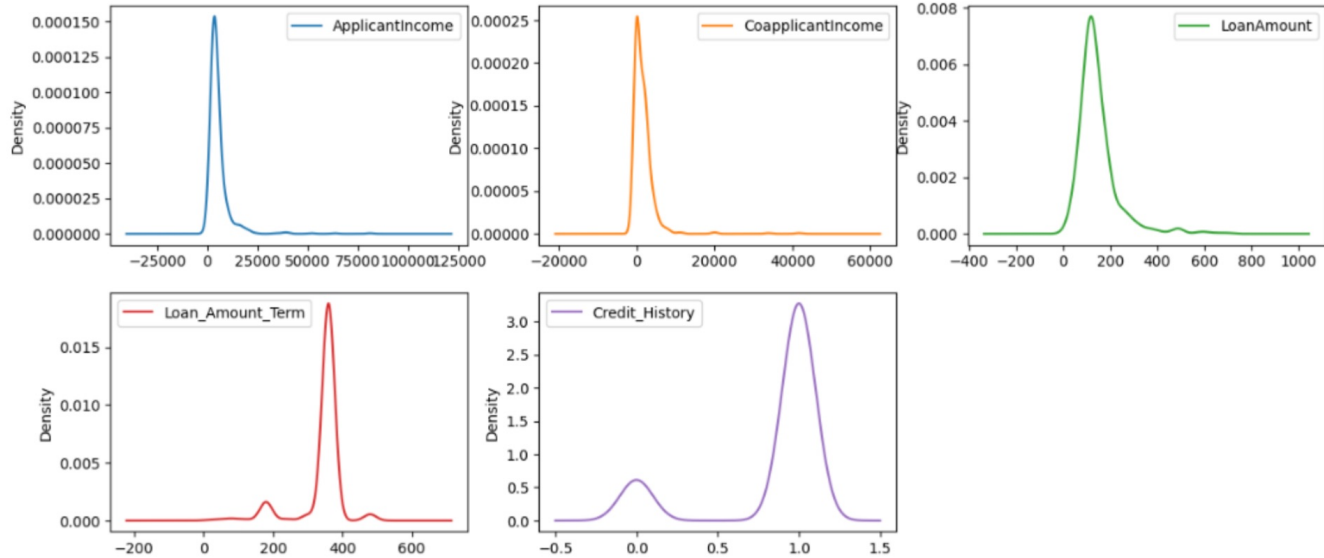
#Skewness

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"

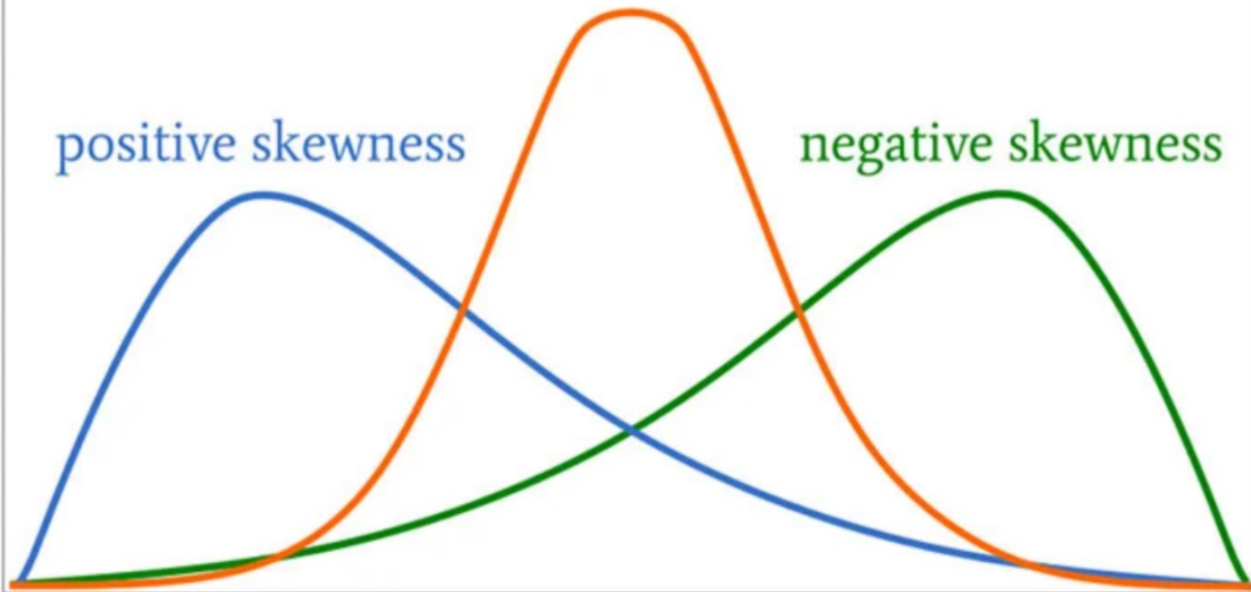
Output



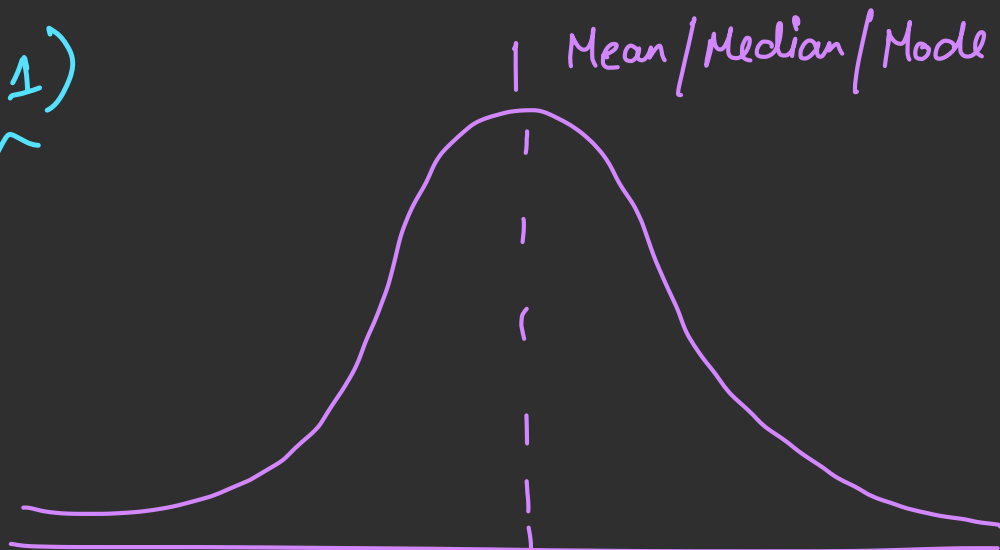
skewness = zero

positive skewness

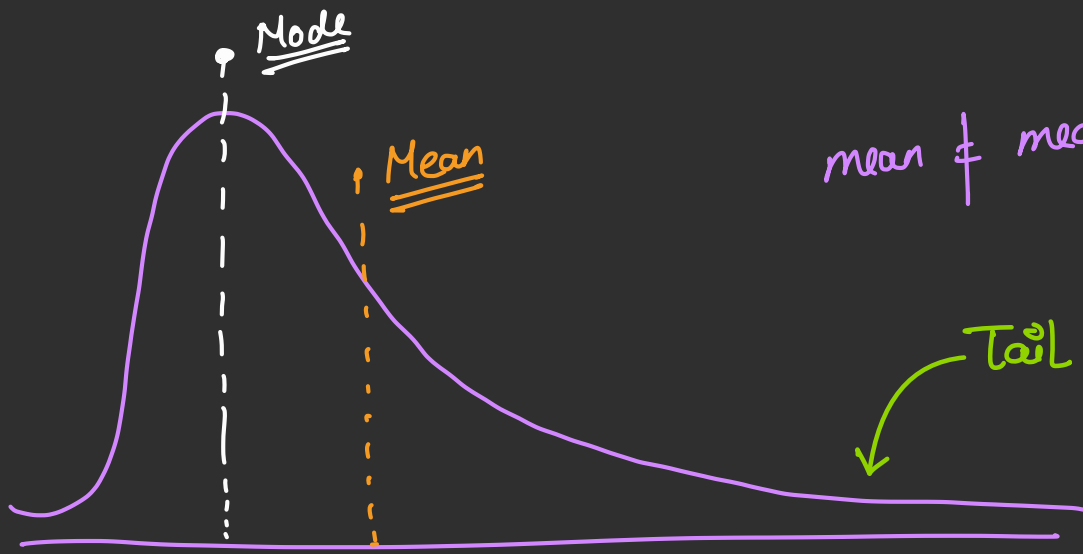
negative skewness



$Z(0,1)$



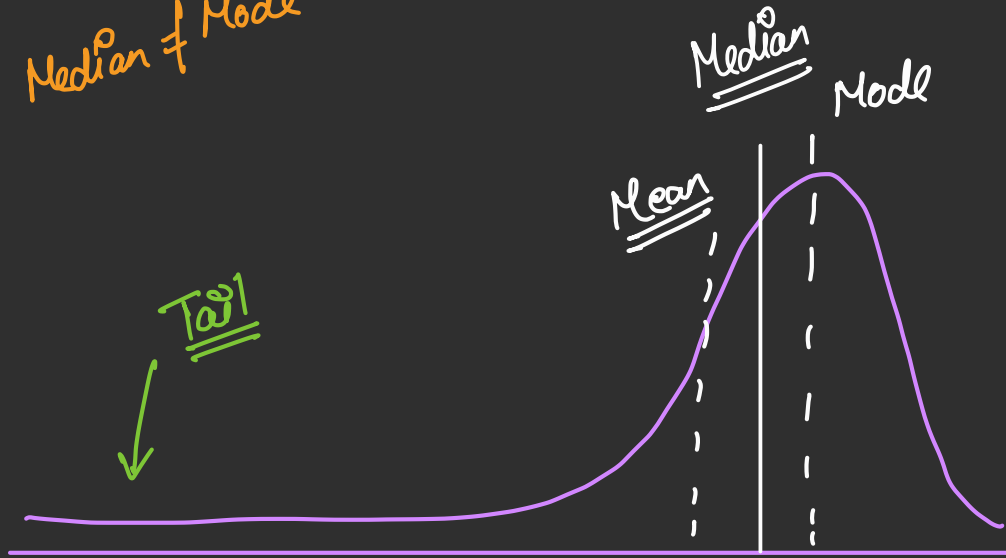
symmetric



mean \neq median \neq mode

unsymmetric
→ (positively / right skewed)

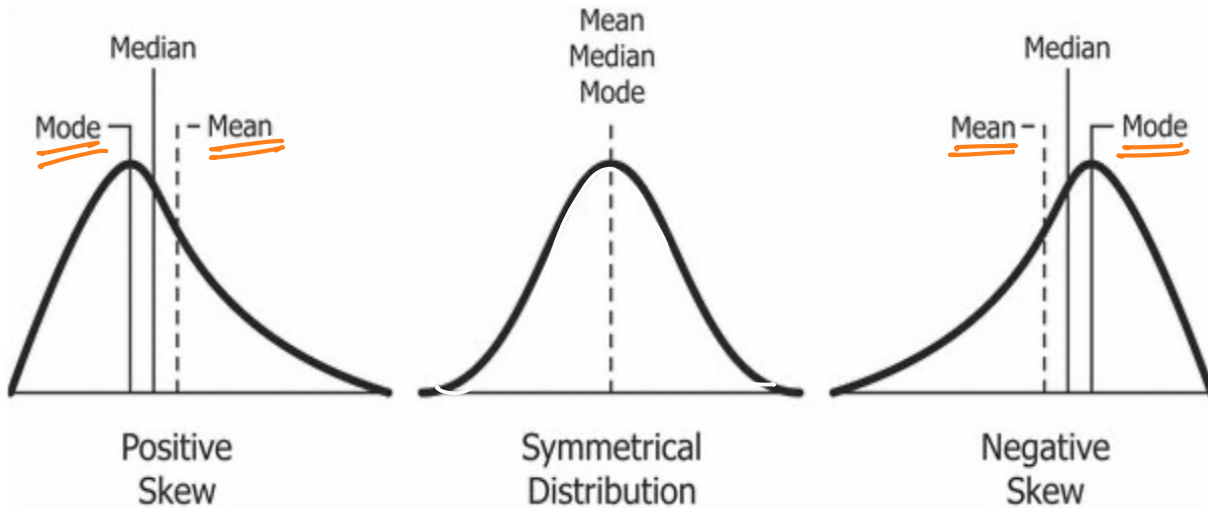
Mean \neq Median \neq Mode



(Neg skewed)

(left skewed)

Types of Skewness



Skewness Formula

✓✓

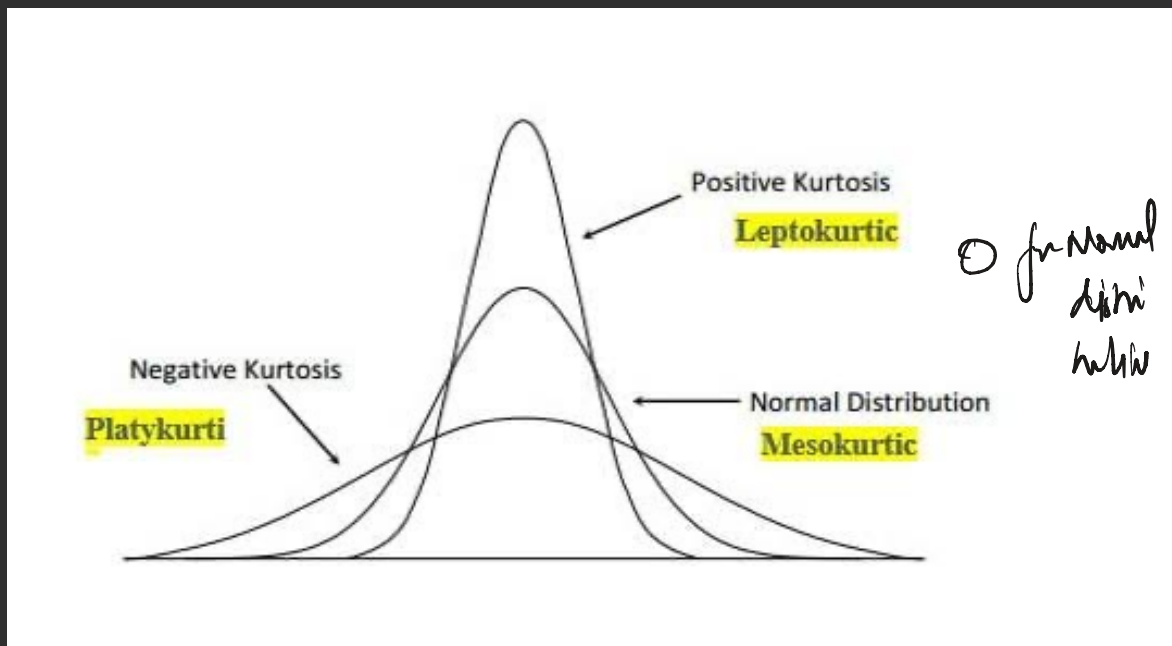
$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \text{mean}}{\text{std dev}} \right)^3$$

→ 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.

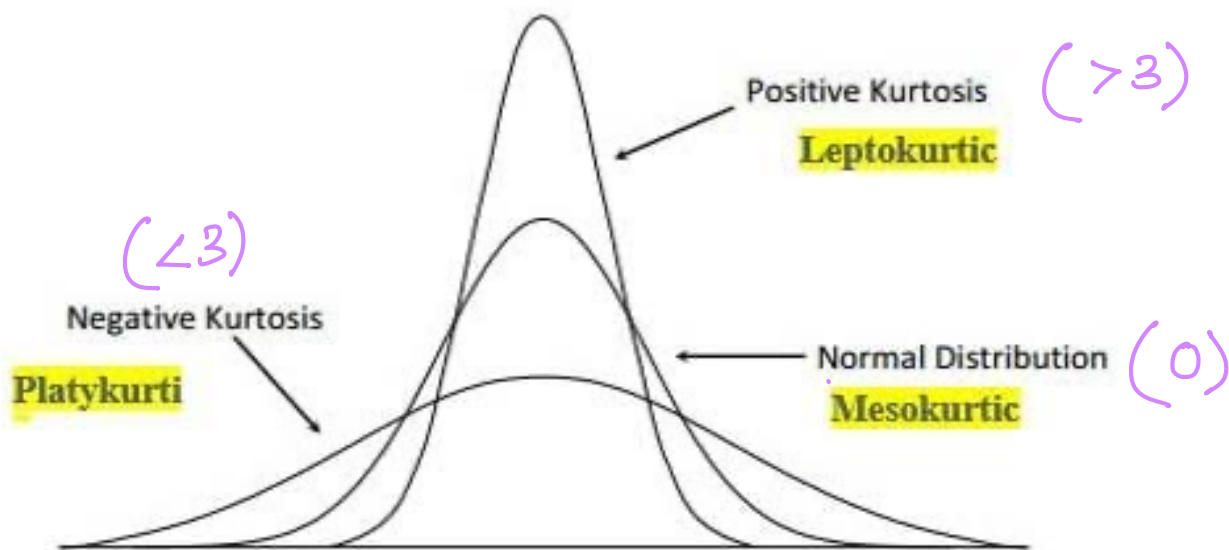


Kurtosis Formula

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

→ 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 \longrightarrow Nature of the dist
- ② Kurtosis \longrightarrow Concentration of data points

#Quiz

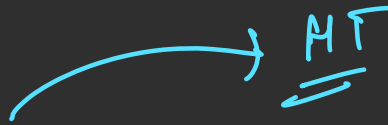
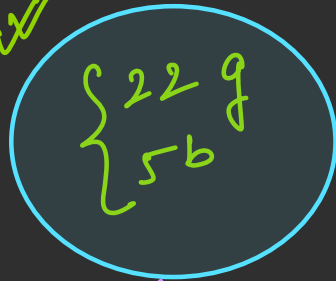
Skewness for Normal Dist $\rightarrow 0$

Kurtosis for Normal Dist $\rightarrow 0$

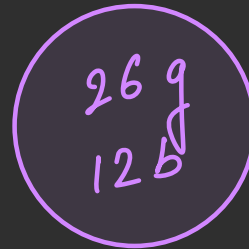
Both are zero for Normal dist

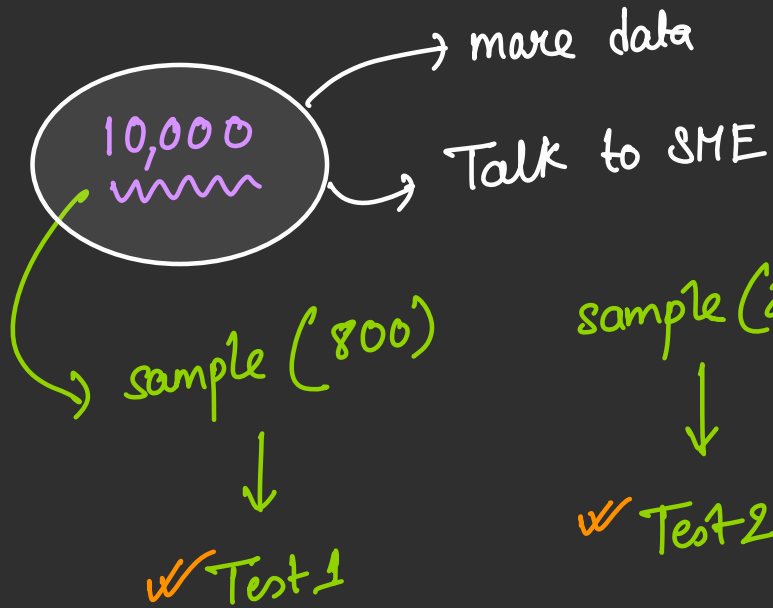
g, b, g, b,

✓✓



MT





sample (800)
↓
✓ Test 2

sample (800)
↓
✓ Test 3

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 \rightarrow shortlisting a Test

1) Create a mental map (broad view)

2) Practice