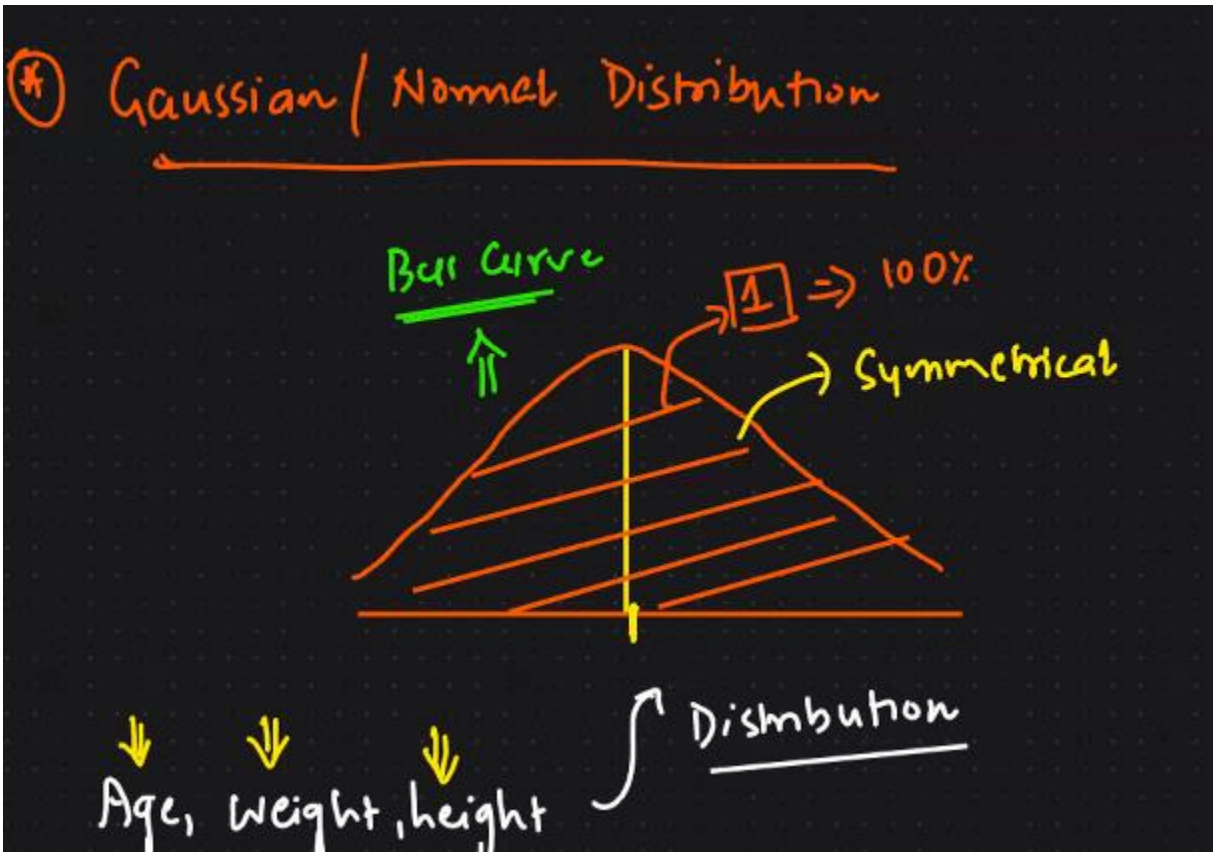
Types of Distributions:

1. Normal distribution:
2. Standard Distribution
3. Z-score
4. Standardization and Normalization

Both the sides of the "Bell curve" is symmetrical and area under the curve lies the data. We create a histogram for Bell curve

Iris dataset and Age , weight, height and most of the dataset follow the **gaussian or normal distribution** as below:



Empirical Rule of Normal/Gaussian Distribution :

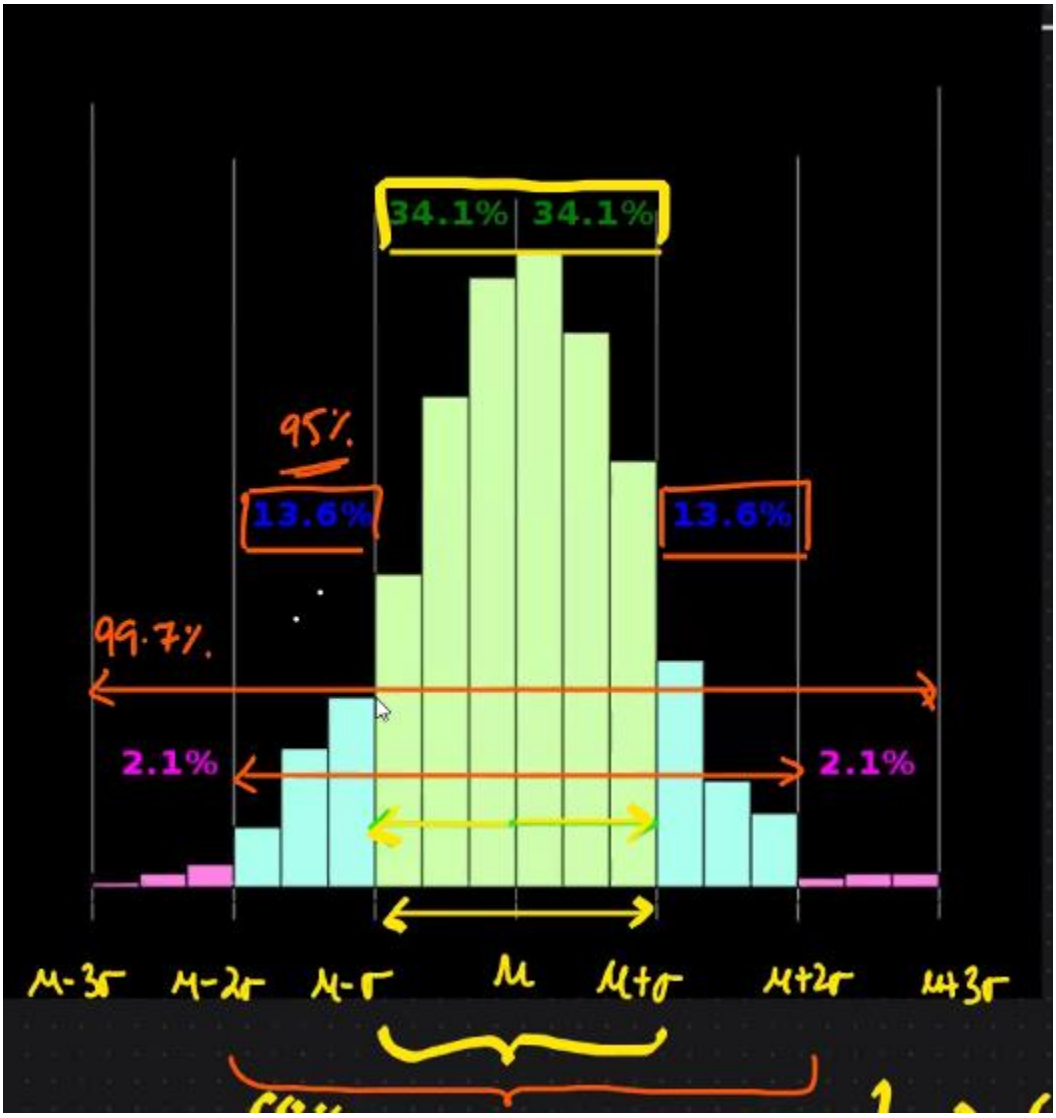
Why Gaussian and Normal distribution plays an important rule for assumption of the data?

Within the **first SD(Standard Deviation)** between the left and right there are around **68%** of data

Within the **2nd SD to the right and left** , there will be around **95% data lies**

Within the **3rd SD of the left and right** around **99.7% data** will lie

This **rule is called as 68-95-99.7%** is called as **Emperical Rule of Normal or gaussian** distribution. Pls see the below screenshot:

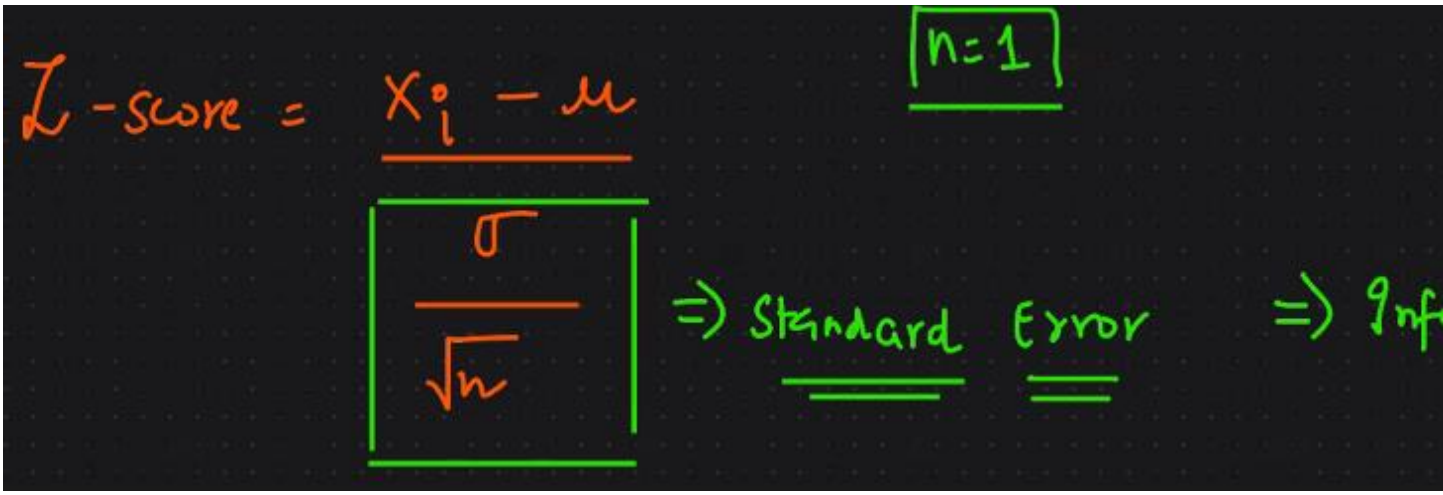


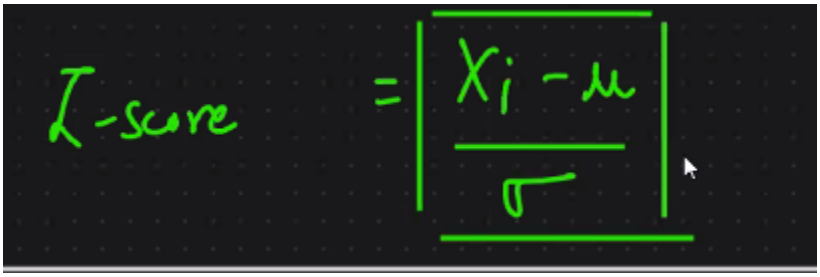
Standard Normal Distribution :

We can transform the gaussian distribution which can be transformed to Stad and distribution with mean = 0 and SD = 1

"n" = Sample Size
The below formula will be applied to each and every element of x

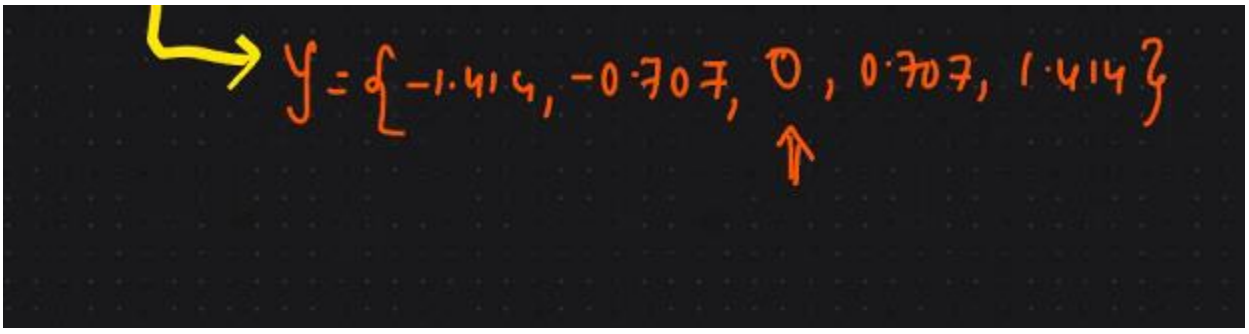
X = {1,2,3,4,5}
Mean = 3
SD = 1.41





So the output will be =
1-3/1.41 = -1.1277 -1.414
2-3/1.41 = -0.707
4-3/1.414 = 0

So now the "y" value =
Y = {-1.414, 0.707, 0,…}



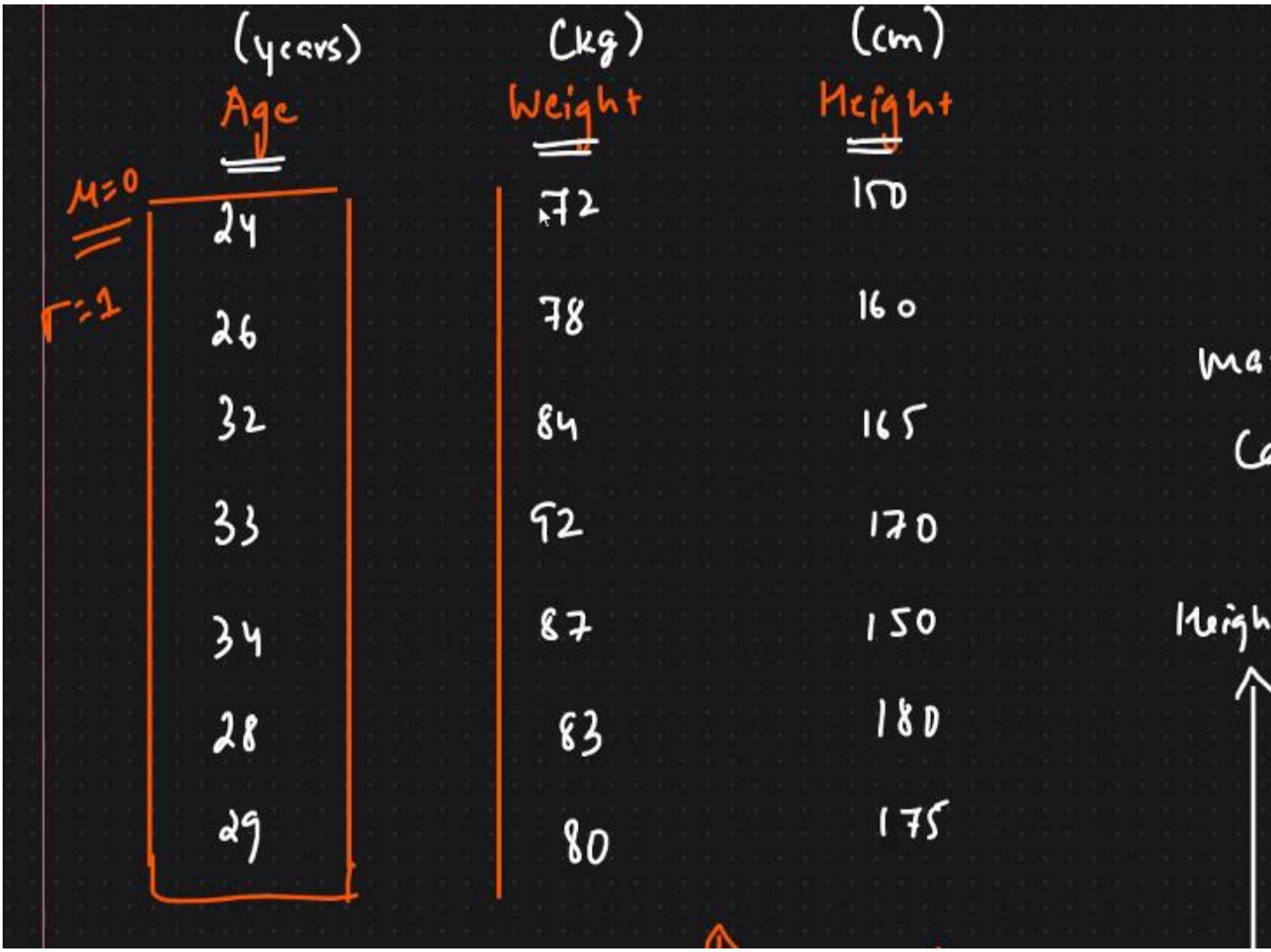
Why are we converting "x" Gaussian to "y" standard normal distribution?

Why?

Age (years)	Weight (kg)	Height (cm)
24	72	150
26	78	160
32	84	165
33	92	170
34	87	150
28	83	180
29	80	175

Mathem

Calcul



For every feature data, we scale down to mean = 0 and SD = 1 by using Z-score

Normalization :

Standardization = We apply z-score here with mean = 0 and SD = 1 so all the values will be converted between +3 to -3

Normalization : In Normalization, we give the range for example :
[o -1]
[0 -5]

Where "0" is a lower scale and "-1 is higher scale
By using a formula called "min max scalar"

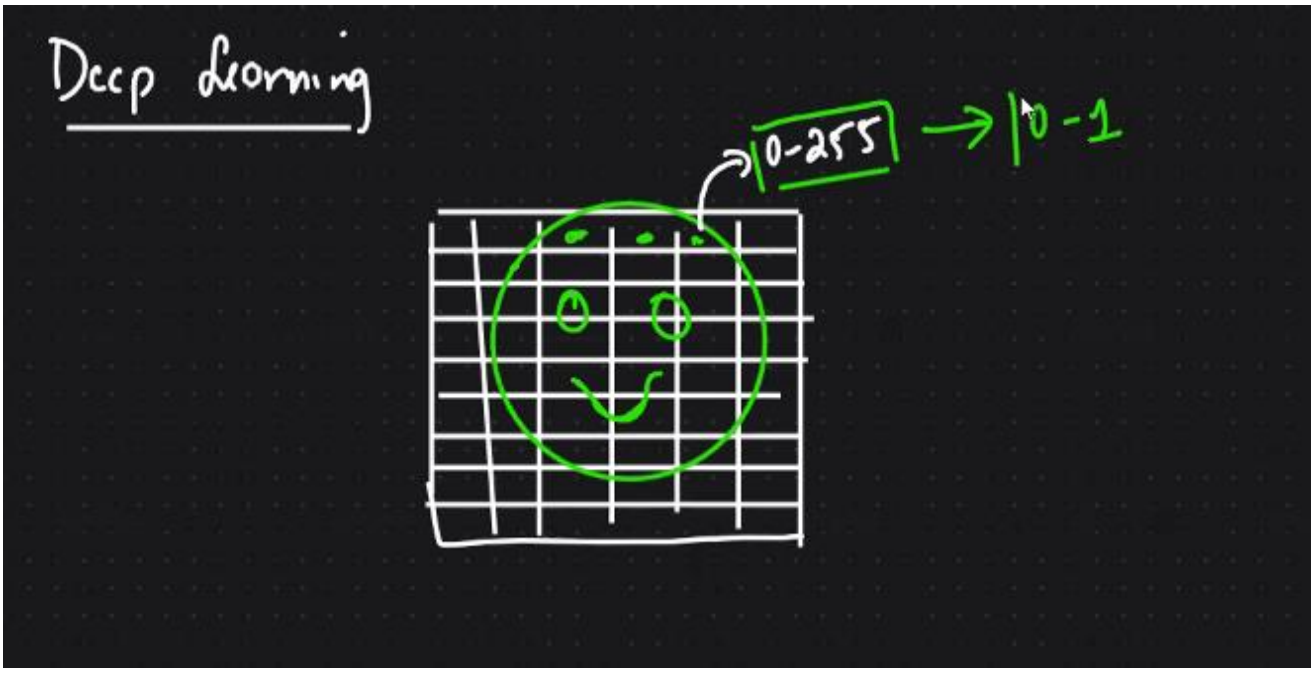
X = [1,2,3,4,5]

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

= 1-1/5-1 = 0
= 2-1/5-1 = 0.8 1/4
= 3-1/5-1 = 1.8 2/4
= 4-1/5-1 = 2.8 3/4

So, with the help of "**Min max scaler"** we **convert the values to "0" to "1".**

We apply "**min max scaler"** in Deep learning and some of the machine learning example " Images pixels and pixels range between **"0" to "255"** and we **convert it** between **"0" and "1"** as shown below:



The "Normalization" and "Standardization" applied in "Feature Scaling Technique"

Which one to use ?

For Machine learning , use Standardization
For Deep Learning CNN, use "Normalization"

Purpose of normalization? :

It is to bring the values into the scale of 0 and 1

Why Z-score used?

Z-score is used to convert "X" which is Normal distribution to "y" using mean = 0 and SD = 1

Why do we use Z-score?

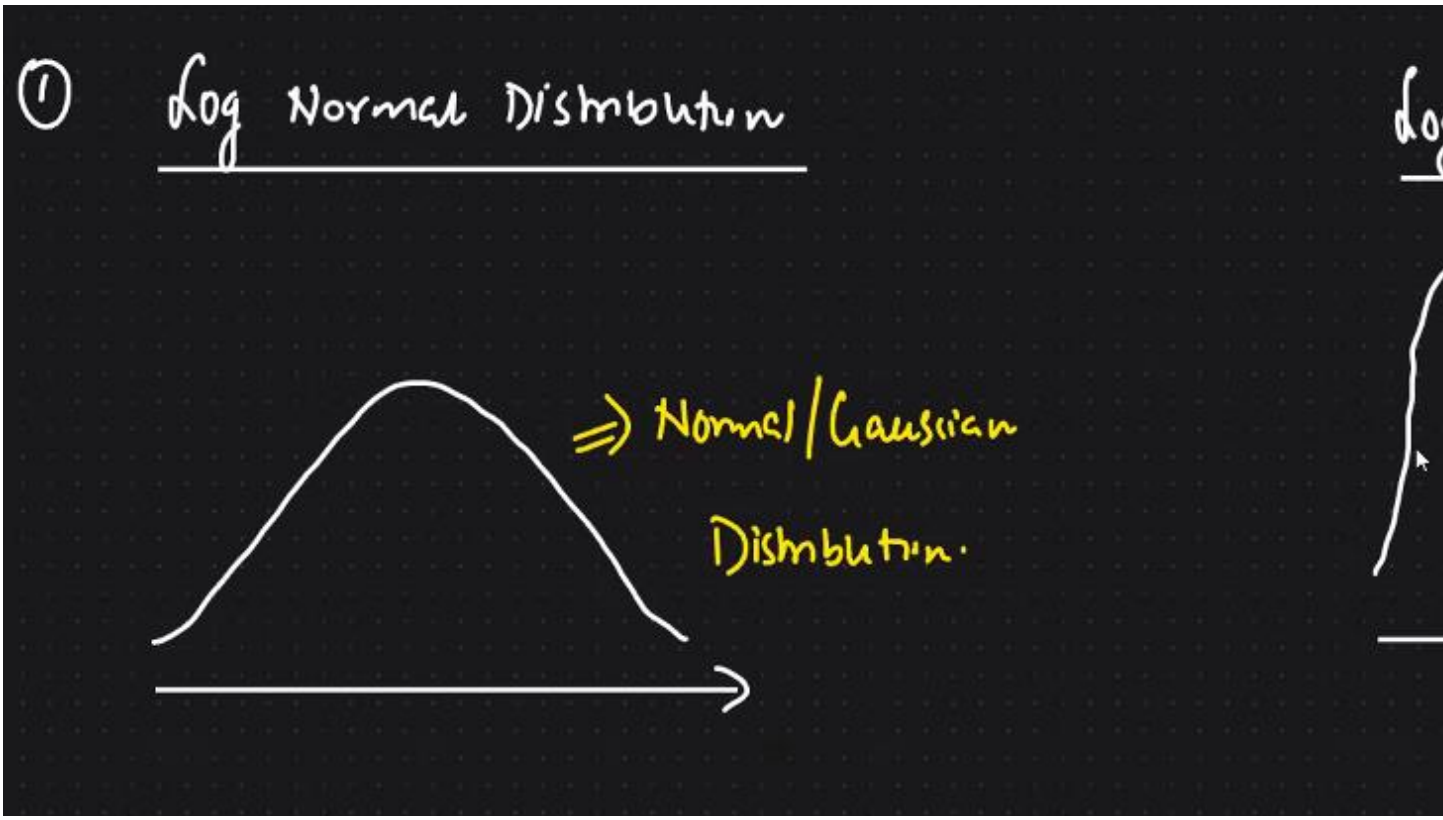
Bring the feature to the same scale.
Time complexity will take less time
Scale down the value and bring all the feature values into the same scale
Calculation will happen quickly and fast
The data will be very scattered

Why do we use Normalization ?

We are giving the range
We use "Min max scaler" to normalize the data between the range we defined

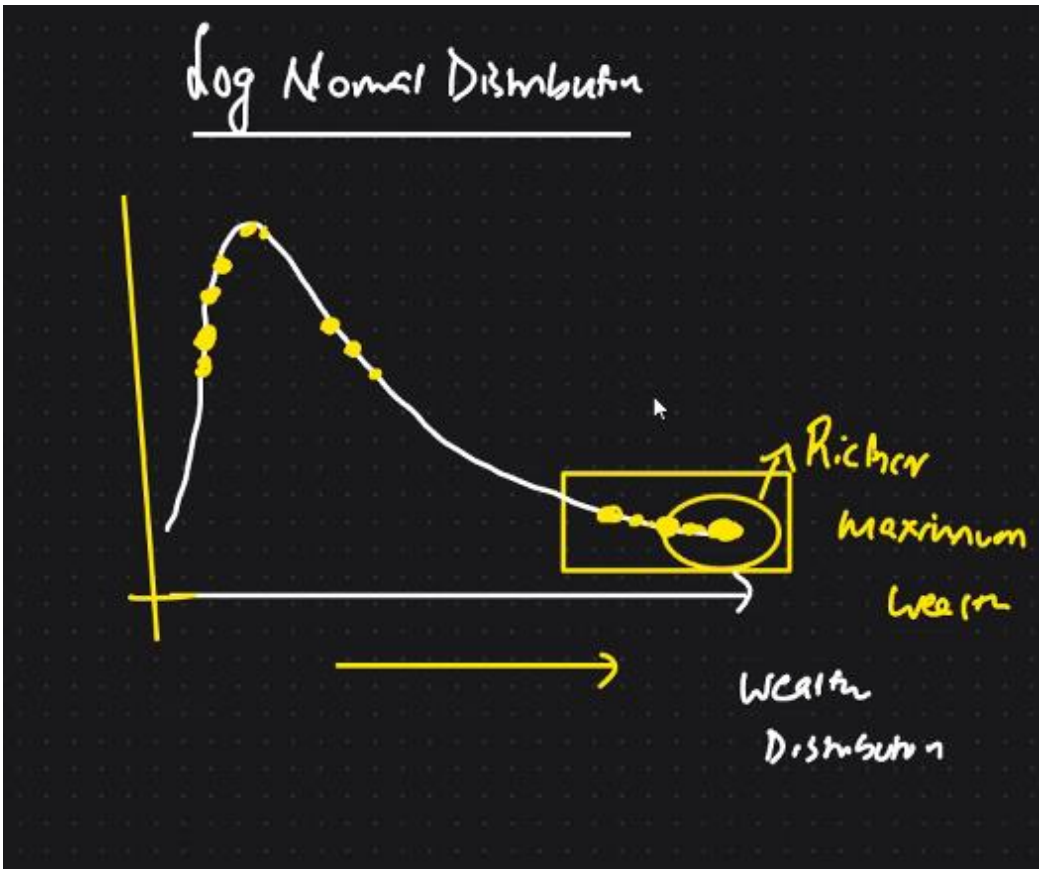
We can use either standardization and Normalization , whichever gives best accuracy we can go with the same

Log Normal Distribution :

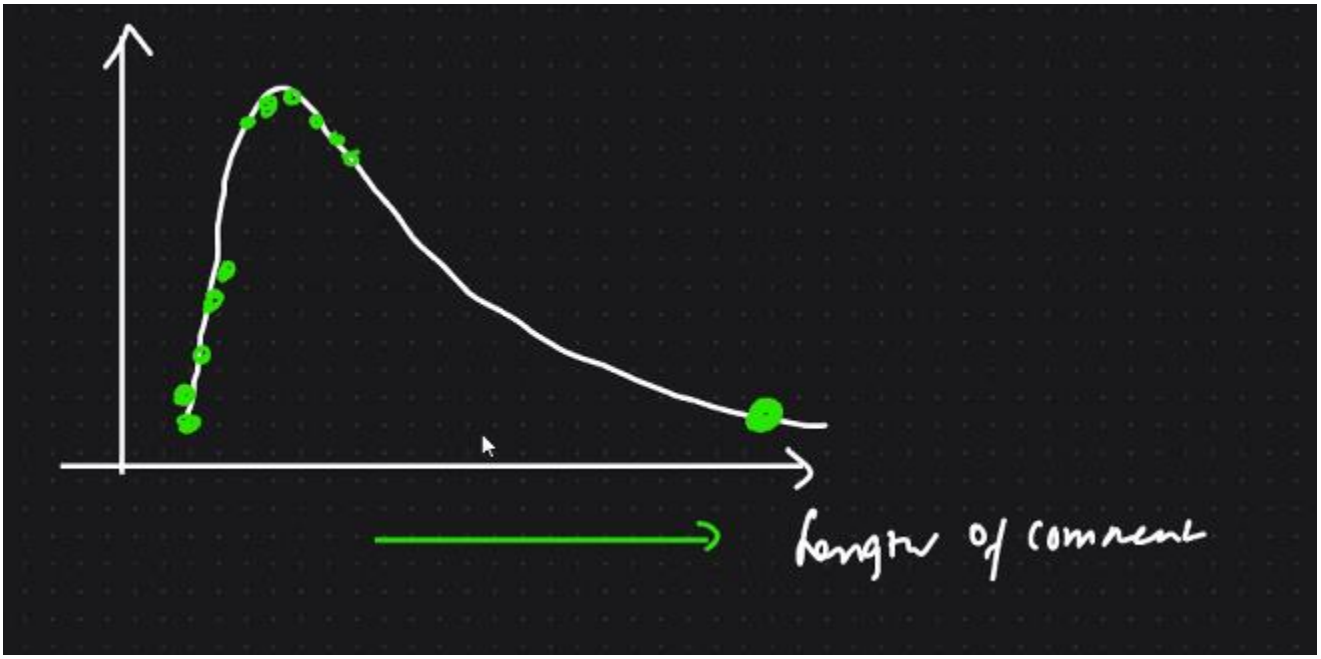


Log normal distribution is skewed on the right side

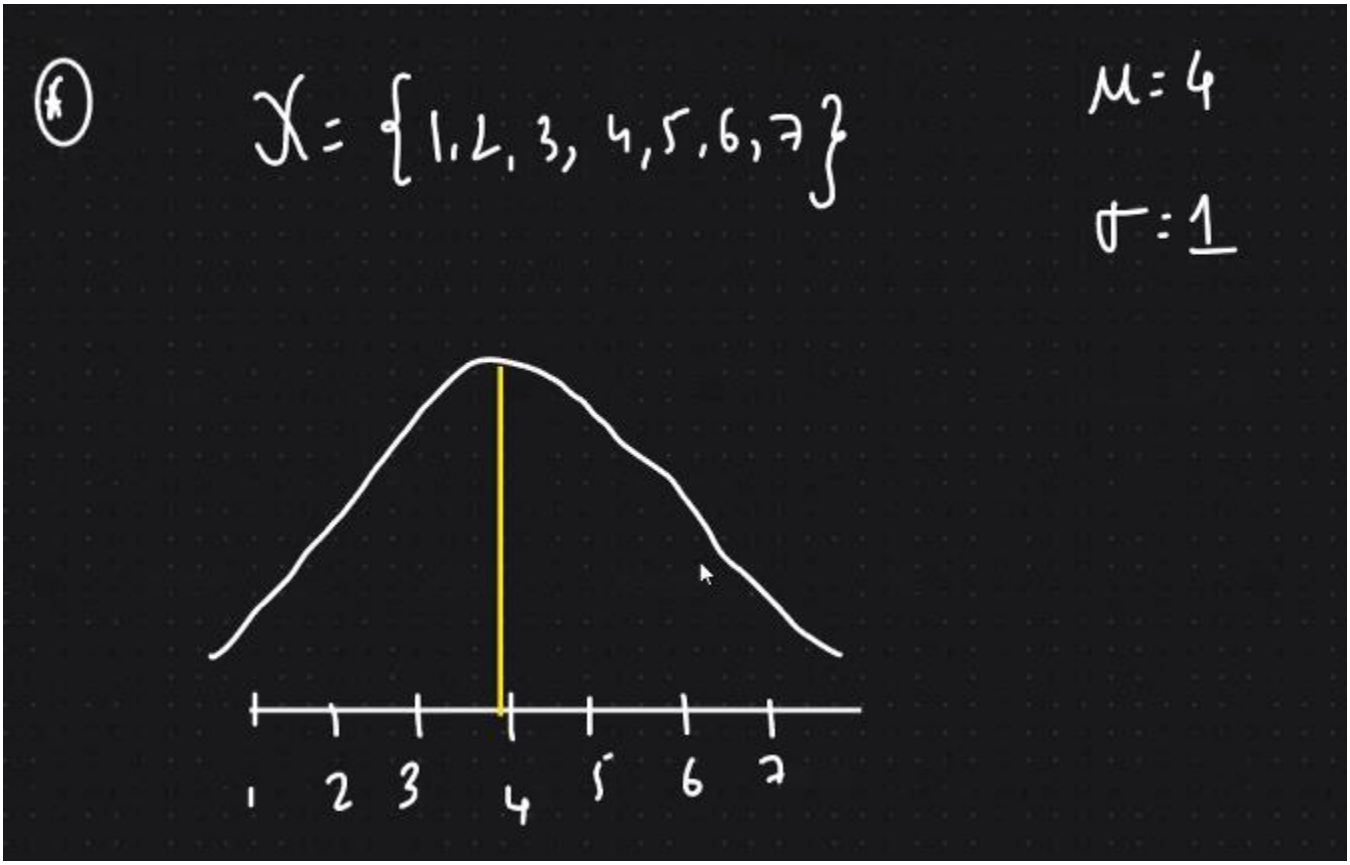
Example : Wealth Distribution



Example : Length of comments in youtube channel:



Problem Statement :



What is the percentage of score that falls above 4.25. Find out the AOC area under the curve ?

Step 1 : AOC of the entire curve is "1"

Step 2 : Apply a Z-score

Step 3 : 4.25 - 4/1 = 0.25

Step 4 : How many SD is away from the mean = 4 is "0.25"

Step 5 : so, now Z-score is "0.25" Standard deviation distance to the righty

Step 6: To find the AOC , area under curve is by using "Z-Table

Step 7: Go to google and search for Z table

Step 8 : In Z table we have negative and positive Z table

Step 9: When we have Negative standard deviation , we will find the area on the left side of the table and whenever we have positive standard deviation

Step 10 : Z-score for "0.25" is ".59" which is 59%

Step 11 : Area above the line is 1-.59 = 0.41 %

Calculate the area between 4.75 and 5.75?

Step 1 : AOC of the entire curve is "1"

Step 2 : Apply a Z-score

Step 3 : 4.75 - 4/1 = 0.75

Step 4 : Z-score of "0.75" is .77

Step 1 : AOC of the entire curve is "1"

Step 2 : Apply a Z-score

Step 3 : 5.75 - 4/1 = 1.75

Step 4 : Z-score of "1.75" is .95

Final Step : Result : -0.18

(4) In India the average IQ is 100 with SD 15. What is the percentage of population have an IQ

(1) Lower than 85

(2) Higher than 85

(3) Between 85 and 100

Assig