Statistics in Data Science

Class # 04

Distributions

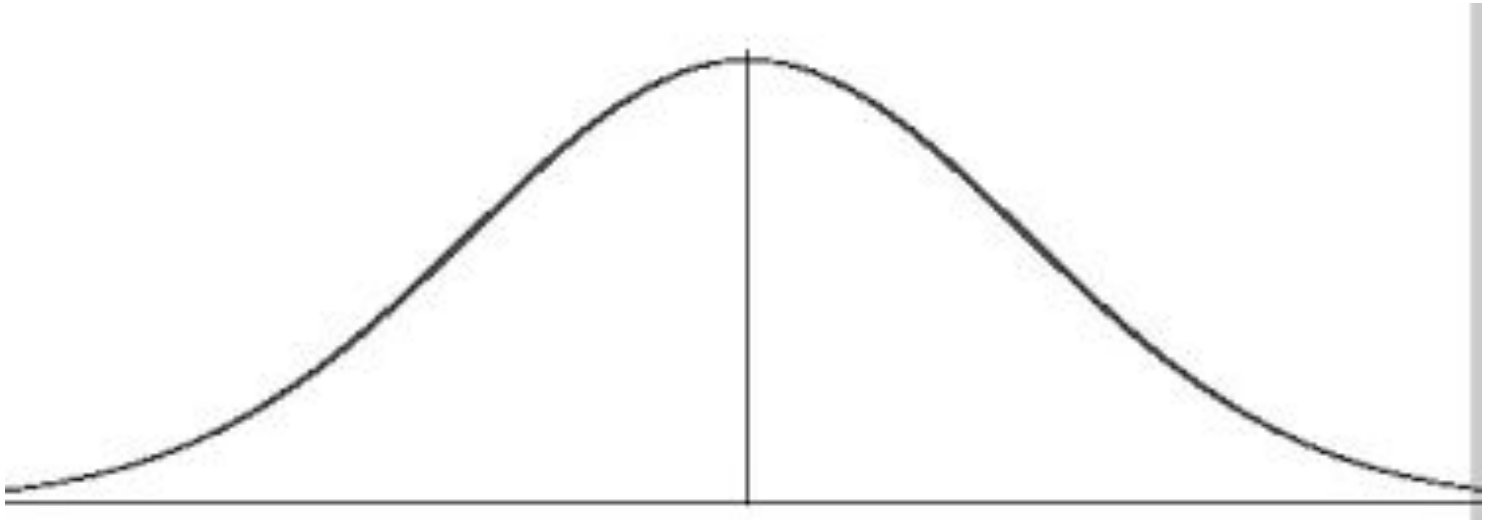
Let's say we have a Data:

Ages = {24, 26 , 27 , 28 , 27 , 22 , 24 , 23 .....}

Now How can we see this Data in a visualize way

There are multiple ways to visualize this data through various graph when we are creating reports

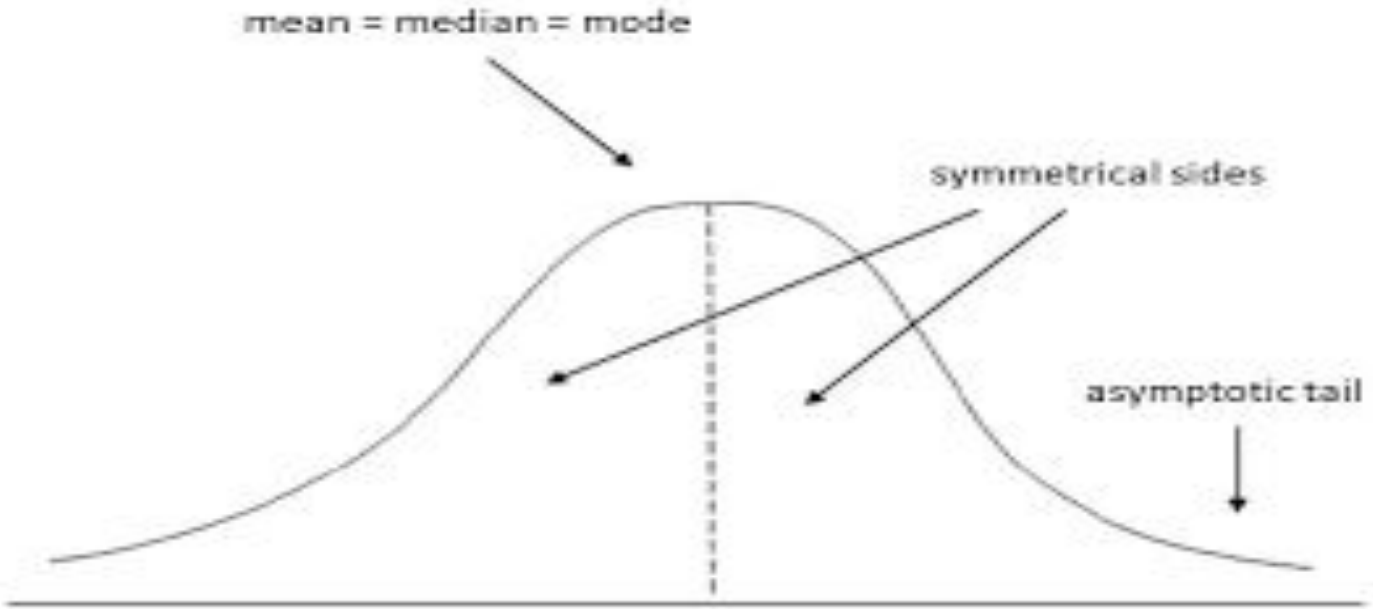
Gaussian Distribution Graph



Continue...

First is why we need Different kinds of Distributions because through this we can have some idea about dataset

So when we talk about Gaussian Distribution we see a “BELL CURVE”

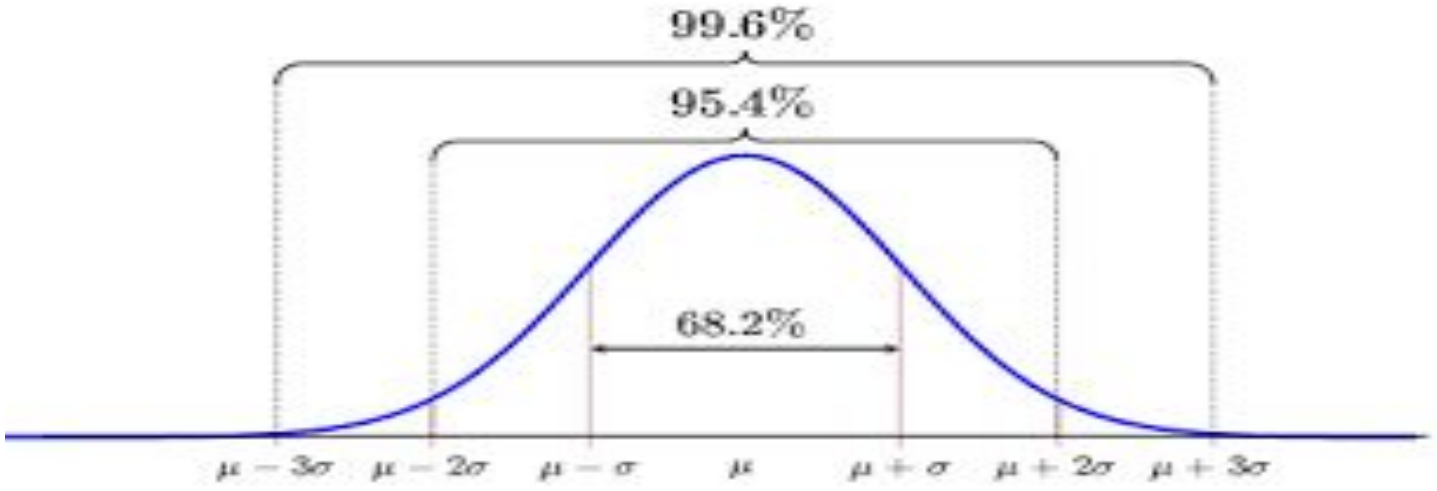

mean = median = mode
symmetrical sides
asymptotic tail

Continue...

Whenever we see this bell Curve the first thing we notices is the left part is exactly symmetrical to the right part that means the amount of Data Present in the right part is exactly same as present in the left part .

Through this distribution we can derive multiple conclusions.

Visualization



Empirical Formula

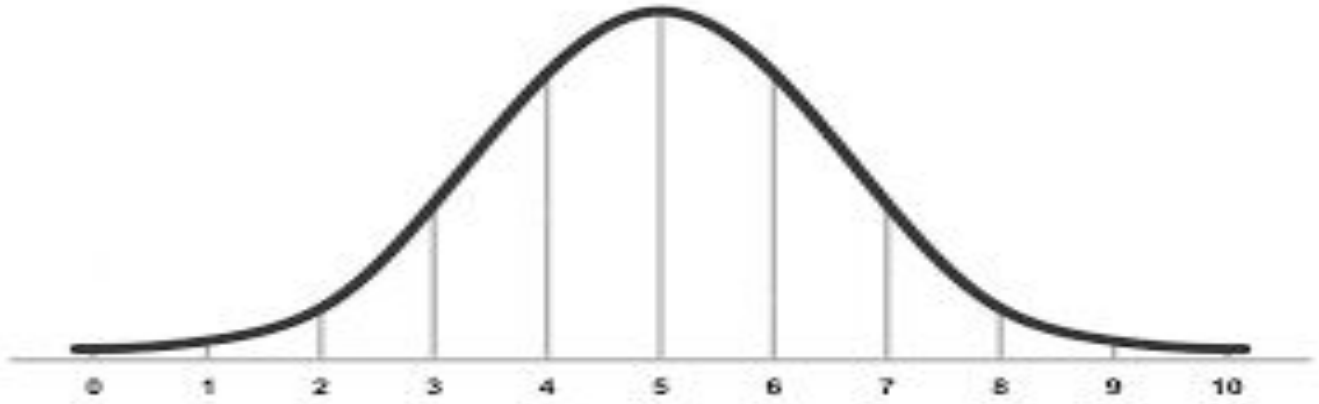
Within the first Standard deviation 68% of data will be present within the 2nd standard deviation 95% of data will present and within 3rd standard deviation 99.7% of data will be present.

This is called empirical Formula

68-95-99.7% Rule

So any Dataset which follows Gaussian Distribution that definitely follow this Rule

Suppose we have a Dataset that mean = 5 and std = 1



Z-score

Now If I ask you where is 5.5 standard deviation? This is easy but if I ask you you where is 4.75 std falls?

4.75 std calculation is difficult

So for this we can use the concept of Z-score

Continue...

Z-Score will basically help you to find how much std away from the mean

Z-Score = (xi -mu)/std

Mean = 5
Std = 1

Z-score = (xi-mu)/std

5.75 - 5 / 1

0.75 /1

0.75

So we can say that 5.75 is 0.75 toward the right from the mean. Why it is right? Because it is positive value.

Find out about 4.75

...