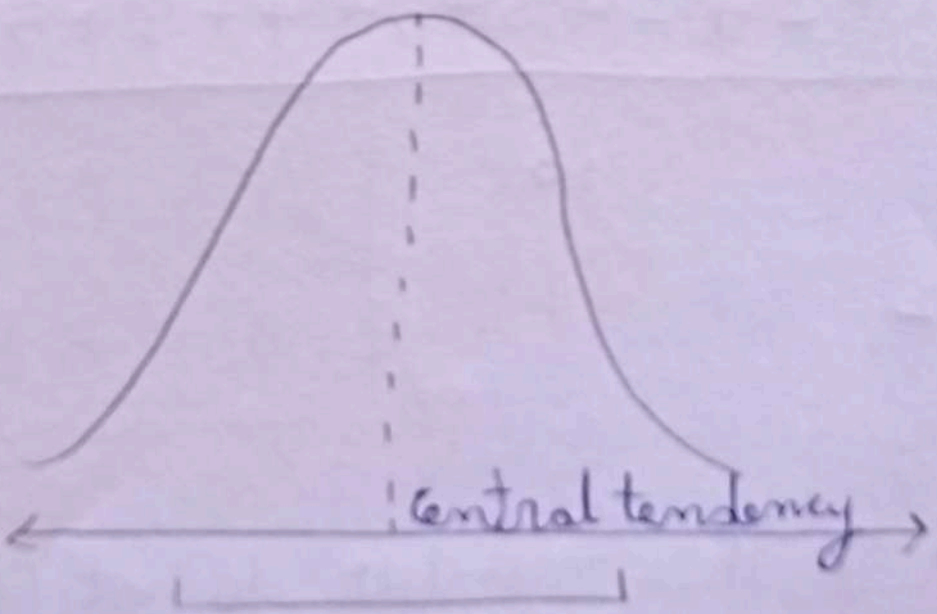
Measures of Central tendency:

① Mean
② Median
③ Mode

[used in EDA & Feature Engineering]

Purpose: Lets say our data is distributed in this manner



maximum amount of data

① Mean

Population (N) Sample (n)

$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

Population mean $(\mu) = \dfrac{\sum_{i=1}^{N} X_i}{N}$

Sample mean $(\bar{x}) = \dfrac{\sum_{i=1}^{n} X_i}{n}$

$\mu = \dfrac{1+1+2+2+3+3+4+5+5+6}{10}$

$= 3.2$

Lets assume X is Sample data.

$n = N = 10$

$\bar{x} = \dfrac{1+1+2+2+3+3+4+5+5+6}{10}$

$= 3.2$

outlier → This number do not belong to distribution. It is very unique number which odd one out of data.

② Median

4, 5, 2, 3, 2, 1

Step 1 → Sort

1, 2, 2, 3, 4, 5

Step 2 → no. of elements

even	odd
1, 2, $\boxed{2, 3}$, 4, 5	1, 2, 2, $\boxed{3}$, 4, 5, 7
$\dfrac{2+3}{2} = 2.5$	3

Why median?

Example 1 →

Sample data = {1, 2, 3, 4, 5}

$\bar{x} = \dfrac{1+2+3+4+5}{5}$ median = 3

$= 3$

Example 2 → If there is an outlier

Sample data = {1, 2, 3, 4, 5, 100}

$\bar{x} = \dfrac{1+2+3+4+5+100}{6}$

$= 19.16$

median $= \dfrac{3+4}{2}$

$= 3.5$

Whenever we have an outlier, median is more relevant

(3) Mode

Element which has maximum frequency in data.

$\{2, \underbrace{1, 1, 1}, 4, 5, 7, 8, 9, 10\}$

Mode = 1

Example ↓

Flower & Age

Flower	Age
Lily	10
Rose	3
______	5
Sunflower	______
Rose	8

Mode is used to fill missing values in a categorical column