1) Measure of Centeral Tendencies :-

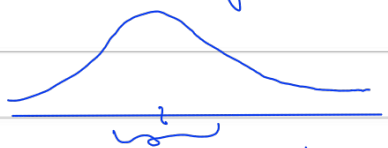
Here we discuss about mainly 3 main topics :-

1. Mean
2. Median
3. Mode

} # EDA (Explonitary Data Analysis) and Feature Engineering will be using these concepts

Q) What does measuring of Centeral Tendency mean?

Soln → Suppose below is the given data distribution :-



So, measure of centeral tendency talks about this centeral region where maximum amount of data is present

1.1) Mean :-

Suppose we have a population of 10 ppl and sample is taken from 5 ppl the calculating both population and sample mean :-

Population ($N = 10$)	Sample ($n = 5$)

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

$\therefore$ Population mean $(\mu) = \frac{\sum_{i=1}^{N} x_i}{N}$

$\therefore$ Sample mean $(\mu_{in}) = \frac{\sum_{i=1}^{n} x_i}{n}$

$\Rightarrow \mu = \frac{32}{10} = \boxed{3.2} = \mu_N$

1.2) Median :-

It is the centeral Element of the sorted list of data

eg: data = 4, 5, 2, 3, 2, 1

⇓

sorting data :- 1, 2, 2, 3, 4, 5

⇓

Median: Count the the no. of Elements :-

if

a) Even Count :- then median is the mean of the two middle nos.

ex: 1, 2, [2, 3], 4, 5.

median = $\frac{2+3}{2} = 2.5$

b) Odd Count :- median is $(n+1)^{th}$ element

ex: - for 1, 2, [3], 4, 4

median = 2

☆5

Q) Why do we need to calculate median?

Soln :- Suppose we have a sample data {1, 2, 3, 4, 5}

mean = $\frac{1+2+3+4+5}{5} = 3$, median = 3

Now,

Suppose an outlayer (100) is added to the above data set such that the dataset becomes

{1, 2, 3, 4, 5, 100}

an outlyer is a value that does not belong/fit into the dataset

∴,

median = $\frac{1+2+3+4+5+100}{6} = 19.16$, mode = $\frac{3+4}{2} = 3.5$

Observation :- Even after adding a large outlayer the value of median have not shifted that much but there is a significant shif in the value of mean

So, whenever we have an outlayer it is a better ~~Strategy~~ to use median cuz change in median value is not significant as compared to the median.

1.34 Mode :-

Element with the maximum frequency will have the mode value

Ex: → Jon {2, 1, 4, 5, 1, 7, 8, 1, 9, 1, 10}

mode = 1

:) Where mode is used?

=> Suppse dataset :-

Type of Flowers	Age
Lilly	10
Rose	3
Rose	5
Sonflower	Median or Mean
Rose	8

Replaced by 'mode'

If outlyer is present the replace with Median or Mean

EDA on Feature Engineering