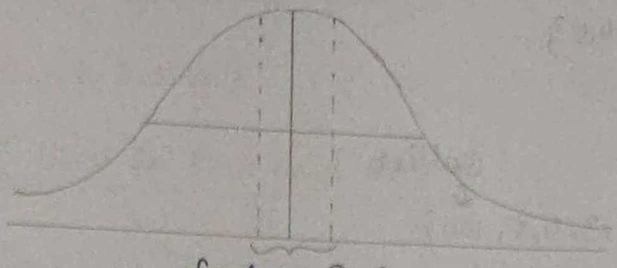
2) Measure of Central Tendency : [EDA]

1. Mean or Average
2. Median
3. Mode



If I want to know what is central data information use one of the above measures.

[Measure of central tendancy are statistical measure used to describe the centre or average of data set. There are 3 measures of central tendancy. These measures help to understand the distribution of data. The choice of which measure to use depends on the type of data you are dealing with.]

---

1. Mean

Population (N)

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

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

$= \frac{1+1+2+2 \dots +5+6}{10}$

$= \frac{32}{10}$

$= 3.2$

Sample (n)

$x = \{2,3,1,4,8,9,5\}$

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

$= \frac{32}{7}$

$= 4.57$

---

2. Median

$x = \{4,5,2,3,2,1\}$

1] Sort the random variable

2] No. of elements → If n = even ⇒ Take avg. of middle 2 elts.
→ If n = odd ⇒ Central element.

$\therefore x = \{1,2,\boxed{2,3},4,5\}$

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

Why Median if you have Mean ?

Let, consider {1,2,3,4,5}

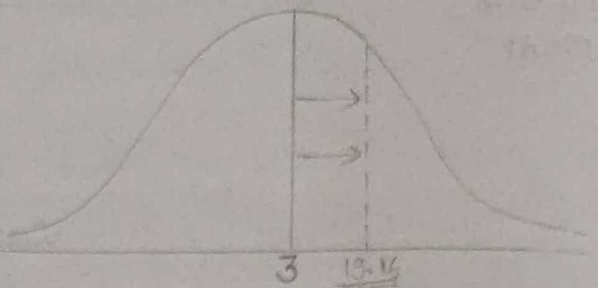
Mean = 3

Median = 3

Now, Consider $\{1,2,3,4,5,\overset{\text{Outlier}}{100}\}$

Mean = 19.16

Median= 3.5



Initially, Mean was 3 and because of the Outlier it is shifted to 19.16 So, if we used Median, there is hardly any minor shift from 3 to 3.5.

So, if you have Outlier, the best thing to use is Median.

- Example of Outlier : Transaction fraudlent
- Example:

Let, $x = \{2,8,4,5,1,7,9,120,130\}$

Find Mean and Median

∴ $\{1,2,4,5,7,8,9,\underline{120,130}\}$

$$\text{Mean} = \frac{1+2+4+5+\dots+130}{9}$$

$$\boxed{\begin{aligned}\therefore \text{Mean} &= 31.77\\ \text{Median} &= 7\end{aligned}}$$

By removing Outlier we have, {1,2,4,5,7,8,9}

$$\boxed{\begin{aligned}\therefore \text{Mean} &= 5.1\\ \text{Median} &= 5\end{aligned}}$$

Here, we can see 2 Outliers and that's why Mean is shifted.

Because of Outlier entire Measure of Central tendancy is moving in case of Mean & In case of Median only a little movement is there.

NOTE : Have Outlier ⇒ Use Median ⇒ To Calculate Central Tendancy

- Example: $x = \{-5,-10,1,2,3,4,5\}$ Mean = 0

⇒ $\{-10,-5,1,2,3,4,5\}$ Median = 2

After removing Outlier ⇒ {1,2,3,4,5}   Mean = 3   Median = 3

3) <u>Mode</u>

Frequency of Maximum occuring element.

Ex: $x = \{2, \underbrace{1,1,1}, 4, 5, 7, 8, 9, 9, 10\}$

∴ Mode = 1 [Used in EDA and FE]

Categorical variable

	Age	Weight	Salary		Gender	Degree
	24	70	40k		M	B.E.
	25	80	70k		F	–
	27	95	45k		F	–
	24	–	50k		M	Phd.
	32	–	60k	{Mode}	–	B.E.
Mean {	$\boxed{-}$	60	–		–	MSc
	$\boxed{-}$	65	55k		–	BSc
	40	72	–		M	B.E.
Median	150 (suppose)					

To handle the missing values,

- you can find Mean/average of all the values. you can replace it with mean.

Let's say we have Outlier 150, then use Median.

- In Gender and Degree Columns, use Mode (repeated ones) to replace the missing values.

Mode are use for categorical Variable replacement.
If M and F are similar, use anyone.

- Based on Outlier, we will use Mean / Median.
- We can also use Mode for Numerical variable but, there are very less chances that values will get repeated.

{ Statistics is all about assumptions