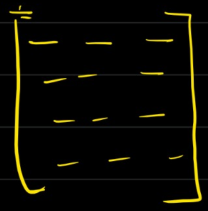
* Descriptive Statistics (summarization of data without adding or subtracting anything at a specific instance/time)

① Measures of Central tendency

② Measures of dispersion

③ Measures of Symmetricity

10:00 AM



① Measures of Central tendency

⇓

Central

1 2 $\overset{\Downarrow}{3}$ 4 5

What is one value around which all the data is revolving??

→ 3

Similarly → who is the one person responsible of your country? Prime Minister.

Representative of your village
⇓
Sarpanch/Mukhya

→ Country
⇓
PM

→ District
↳ M.P.

* Country to respresented by ↑

* CT represents the center point of a dataset

① Mean
② Median
③ Mode

→ Exploratory data Analysis | Data Preparation/ feature Engineering.

① Mean (Average: Arithmatic mid value of data)

Population = {1, 2, 3, 4, 5}

$$\mu = \sum_{i=1}^{n} \frac{x_i}{N}$$

Sample (n)

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Σ → Summation.

$$\frac{\sum_{i=0}^{n-1} x_i}{N} \qquad x = \{\underset{0}{\overset{\uparrow}{1}}, \underset{1}{\overset{\uparrow}{2}}, \underset{2}{\overset{\uparrow}{3}}, \underset{3}{\overset{\uparrow}{4}}, \underset{4}{\overset{\uparrow}{5}}\}$$

$0 - n-1$

or $1 - n$

$$\Rightarrow \frac{\sum_{i=0}^{5-1} x_i}{N} = \frac{x_0 + x_1 + x_2 + x_3 + x_4}{5} = \frac{1+2+3+4+5}{5}$$

$$= \frac{15}{5} = 3$$

* Mean – Summing up all the observations and dividing by no. of observation.

② Median (Physical mid point of data)

4, 5, 2, 3, 1, 2

→ Sort the data – 1, 2, 2, 3, 4, 5.

→ Count the no of elements – 6

if count is even

$\cancel{1}, \cancel{2}, \underbrace{2, 3}, \cancel{4}, \cancel{5}$

median = average of two middle most element

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

4 5 2 3 1

→ Sort : $\cancel{1}, \cancel{2}, 3, \cancel{4}, \cancel{5}$

→ Count – 5

↓

odd

median = the middle most element

= 3

* Physical mid point, (Median: middlemostelement)

1 2 3 4 5

assume the nos are human beings. ⟶ middle most human being :- 3

even if you change 5 to 100.

1 2 3 4 100

$\Uparrow$

Scenerio 1
(mean)

* 1, 2, 3, 4, $\underline{5}$

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

→ 1, 2, 3, 4, 1000

→ Here 1000 is an outlier

⇓

number which is much higher or lower as compared to other numbers (Extreme values)

$$\text{mean} = \frac{1+2+3+4+1000}{5} = \frac{1010}{5} = \underline{202}$$

Scenario 2
(median)

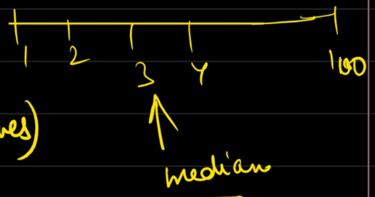
* 1, 2, 3, 4, 5

median: $\cancel{1}\ \cancel{2}\ \underline{3}\ \cancel{4}\ \cancel{5}$

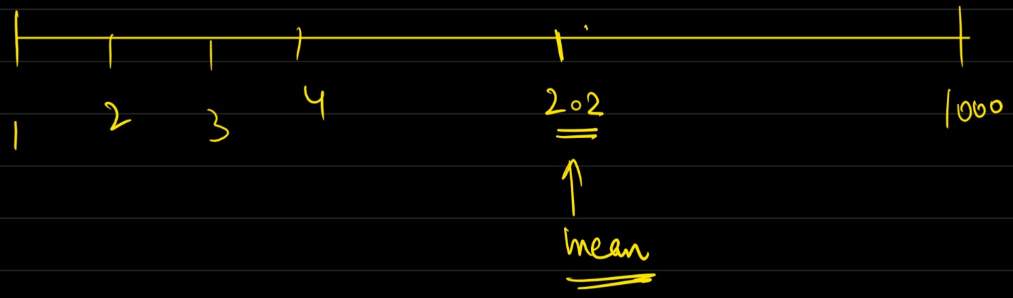
median = 3

$\cancel{1}, \cancel{2}, \underline{3}, \cancel{4}, \cancel{1000}$

median = $\underline{3}$



* The mean is affected (→inflated) by outliers whereas median is not affected by outlier.



③ mode — Maximum frequency (repeated higlest no of times)

{2, 3, 1, 1, 4, 4, 4, 3, 4, 2}

mode = 4

* In all of these cases (mean, median, mode), we are trying to represent the data with a central value.

* On which type of data mean, median, Mode is Calculated ?

date

→ Numerical date (continuous) → mean, median

→ Categorical → Mode.

Use cases of Central tendency (missing value imputation)

	Age	Gender	Weight	Salary (K)
→ null value	25	M	80	—
	26	— → M	70	50
→	— 24.75	M	30	60
	23	M	—	70
	25	F	1000	45

* Age

→ Age is continuous variable

→ Impute|replace the missing|null value with mean if there is no outlier. $\frac{25+26+23+25}{4} = 24.75$

* Gender → Categorical variable

→ Highest frequency

→ M, M, M, F ⇒ Mode = M

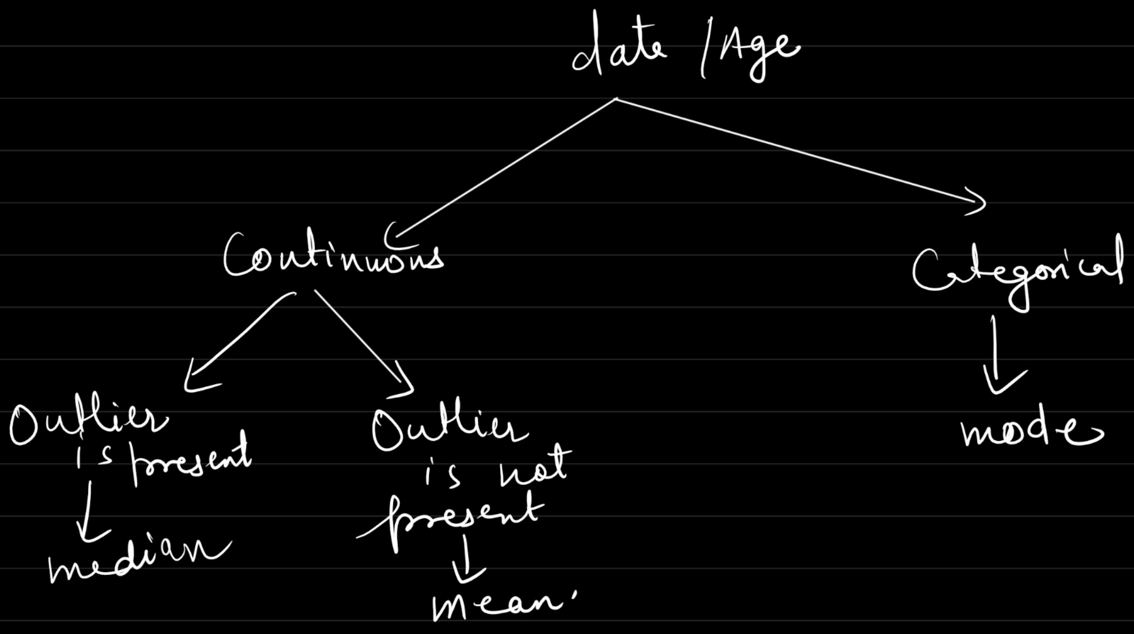
* weight

→ Continuous variable

→ Outlier is present

→ Median will be used for imputation as median is not affected by outlier.

Summarise


date /Age
Continuous
Categorical
Outlier is present
median
Outlier is not present
mean.
mode