* Descriptive Statistic — Summarization of data without adding or subtracting anything at a specific instance.

① Measures of Central tendency

② Measures of dispersion

③ Measures of symmetricity

① Measures of Central tendency

Central → 1, 2, 3, 4, 5

What is one value around which all the data is revolving? — 3

* CT represents the center point of a dataset.

① mean
② median
③ Mode

} EDA and feature Engineering

① Mean (Arithmetic 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 $\Rightarrow \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

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

1, 2, 2, 3, 4, 5

median = avg of two middle most element

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

4, 5, 2, 3, 1

→ Sort – 1, 2, 3, 4, 5

→ count – 5

↓

odd

median = the middle most element

= 2

Scenario 1 (mean)

1, 2, 3, 4, 5

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

Scenario 2.

1, 2, 3, 4, 5

median = 3

1, 2, 3, 4, 1000

→ Here 1000 is an outlier

↓

no which is much higher or lower as compared to other numbers. (Extreme Values)

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

1, 2, (3), 4, 1000

median = 3

* The mean is affected by outliers whereas median is not affected by outliers.

③ mode — frequency maximum

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

mode = 4

Use cases of Central tendency

Age	Gender	Weight	Salary (k)
25	M	80	—
26	— → M	70	50
→ 24.75	M	30	60
23	M	—	~~70~~
25	F	60	45

→ Age is continuous variable

→ impute the missing/null value with mean

$$\frac{25+26+23+25}{4} = 24.75$$

→ Gender → Categorical data

→ Highest frequency

→ M, M, M, F ⇒ Mode — M

* Median is not affected by outlier.