PR-1 Descriptive Booster

Part A: Theory

Q1 Define the following with example from Dataset:
1) Types of Data: Numerical and Categorical
2) Types of Statistics: Descriptive / inferential
3) What is Descriptive Statistics?

Ans 1) Numerical Data: Numerical data measures quantities, all quantitative data which can be measured.
→ Two types of numerical
1) Discrete
2) Continuous

Discrete: It measures countable numbers.
eg: Age (From dataset)

Continuous: It measures all decimal values (points X eg: Heights)

Example: Age_of_household_head, Household_income, family_size

2) Categorical Data: Categorical data are labels which are in qualitative terms
→ Two types of categorical
1) Nominal
2) Ordinal

Nominal: This are those labels which doesn't have any specific order.

eg: Urban_Rural (From dataset)

Ordinal: This are those labels which has an specific order.
eg: education_level (From dataset.

Example: household_id, education_level, Avg_house, Urban_rural.

2) Descriptive Statistics: DS
→ Descriptive Statistics is the branch of statistics that focuses on summarizing raw data in a meaningful way.

Types of Descriptive Statistics:
1) Measures of Central Tendency
   1) Mean
   2) Median
   3) Mode

2) Measures of Central dispersion:
   1) Range
   2) Variance
   3) Standard deviations.

3) Shape of Distribution
   1) Skewness
   2) Kurtosis

4) Data Visualization

Inferential Statistics: Making prediction and inferences based on population for sample.

Types of Inferential Statistics:

1) Hypothesis Testing

2) Confidence Interval

3) Descriptive Statistics: Descriptive Statistics is branch of statistics focuses on summarizing raw data in a meaningful way.

Types of Descriptive Statistics:

1) Measures of Central Tendency

2) Measures of Despersion

3) Shape of Distribution

4) Data Visualization.

Q2 Explain the difference between:
1) Mean, Median, Mode

2) Range, Variance, Standard deviations

Ans 1) Mean: The average value of the dataset

$$\text{Mean} = \frac{\text{Sum of all Values}}{\text{Number of Values}}.$$

Median : The middle value of the dataset.

Odd (Median) = $\frac{n}{2}$ Even (Median) = $\frac{n}{2} + \frac{n-1}{2}$

Mode : The repeated value or the most frequent value in dataset.

2) Range : The difference between the maximum value and minimum value.

Range = Max - Min.

Variance : Variance shows how far each data point is away from mean.

$$\text{Variance} = \frac{\Sigma (x_i - \bar{x})^2}{n}$$

Standard deviation : The Square root of Variance

$$\sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}}$$

Q3 Explain the following terms with neat and clean diagram along with its formula?

Ans 1) Gaussian Distribution : A Gaussian (normal) distribution is a Or bell-shape curve that shows how data is

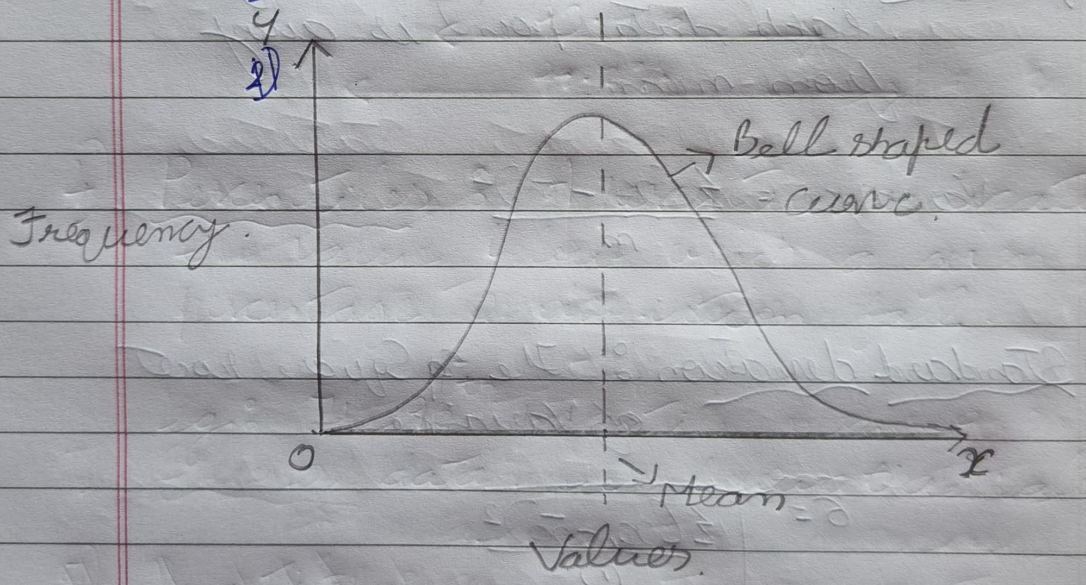
distributed.

Examples:
1) Height of people
2) IQ Scores.
3) Blood pressure.

→ Normal distribution formula:

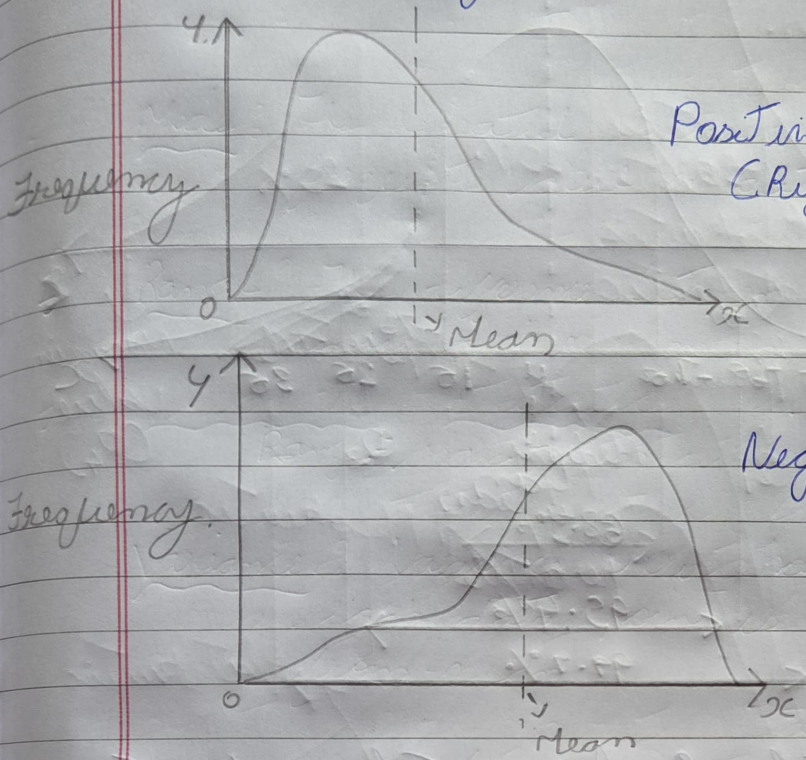
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

→ Properties:
1) Mean ≅ Median ≅ Mode



2) Log Normal Distribution: A log normal distribution is a probability distribution of a random variable whose logarithm is normally distributed.

→ Log normal distribution looks like a positively skewed and Negatively skewed (long tails ⊥ or respective sides).



→ Characteristics:

1) A Log-normal only contains positive values.

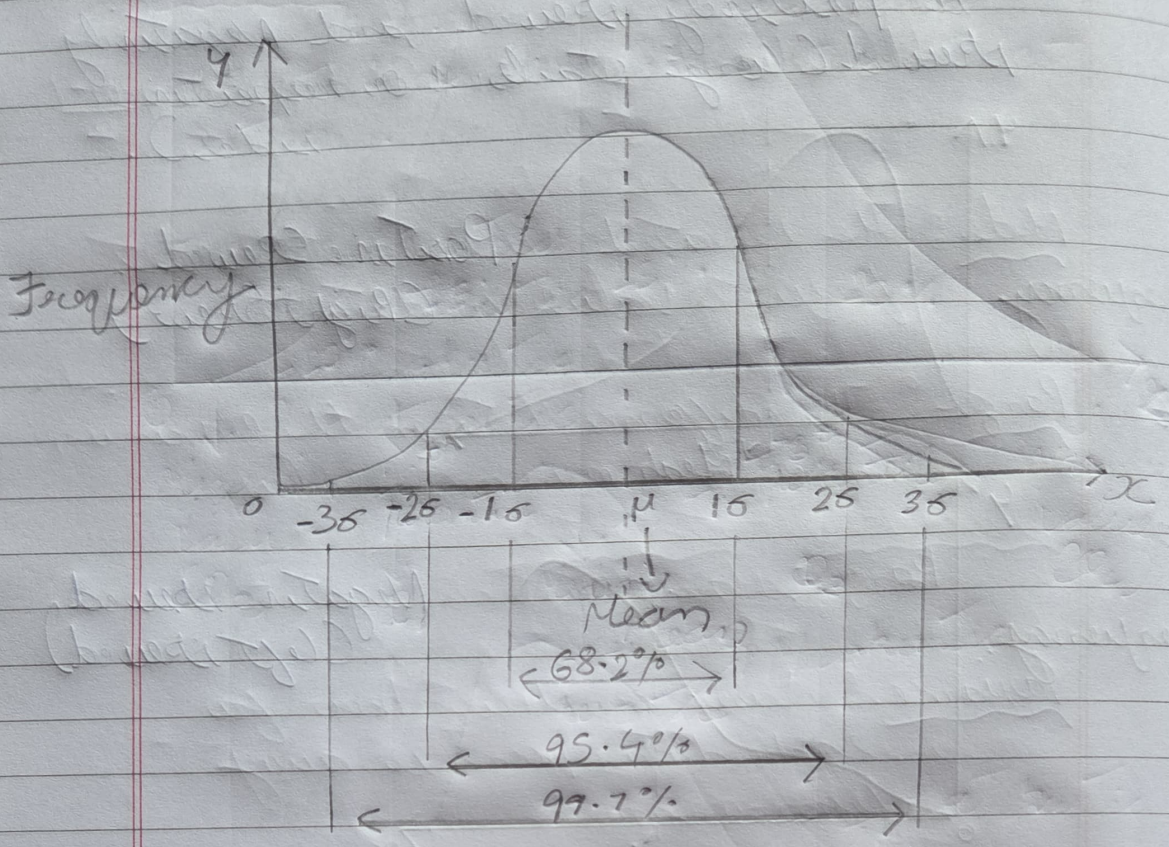
3) B) 3-Sigma Rule or Emperical Rule:

→ Out of 100% How much data lies in which part of standard deviation (σ) is called 3-Sigma-Rule.

1) 68.2% of values lies from -1σ to +1σ (standard deviation).

2) 95.4% of values lies from -2σ to +2σ

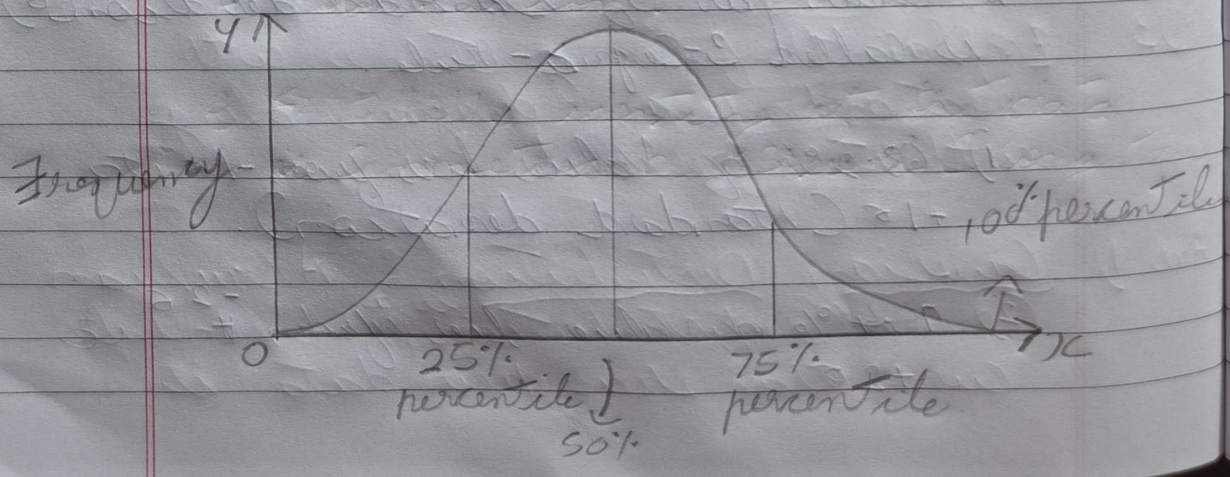
3) 99.7% values lies within -3σ to +3σ



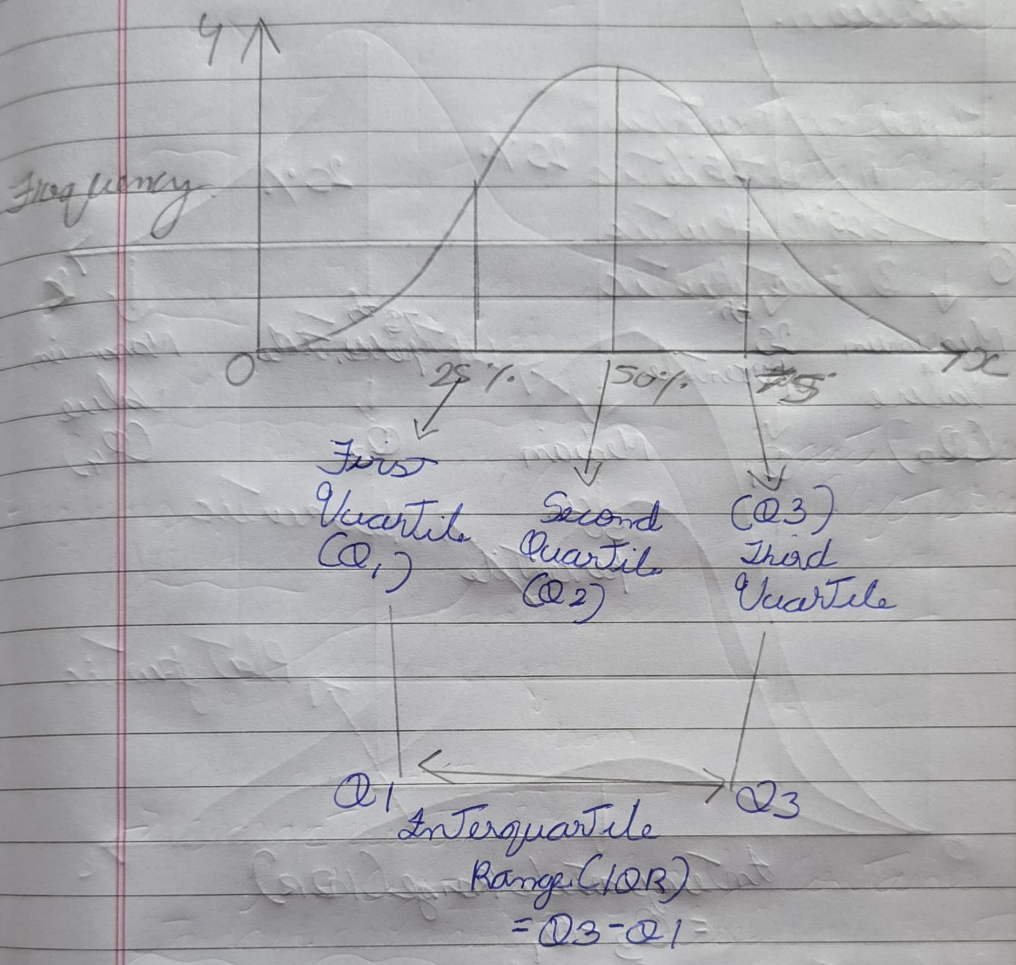
4) Percentiles : A percentile indicates the value below which a given percentage of observation fall.

eg: The 90th percentile means 90% of data lies below that value.

5) Quartiles :

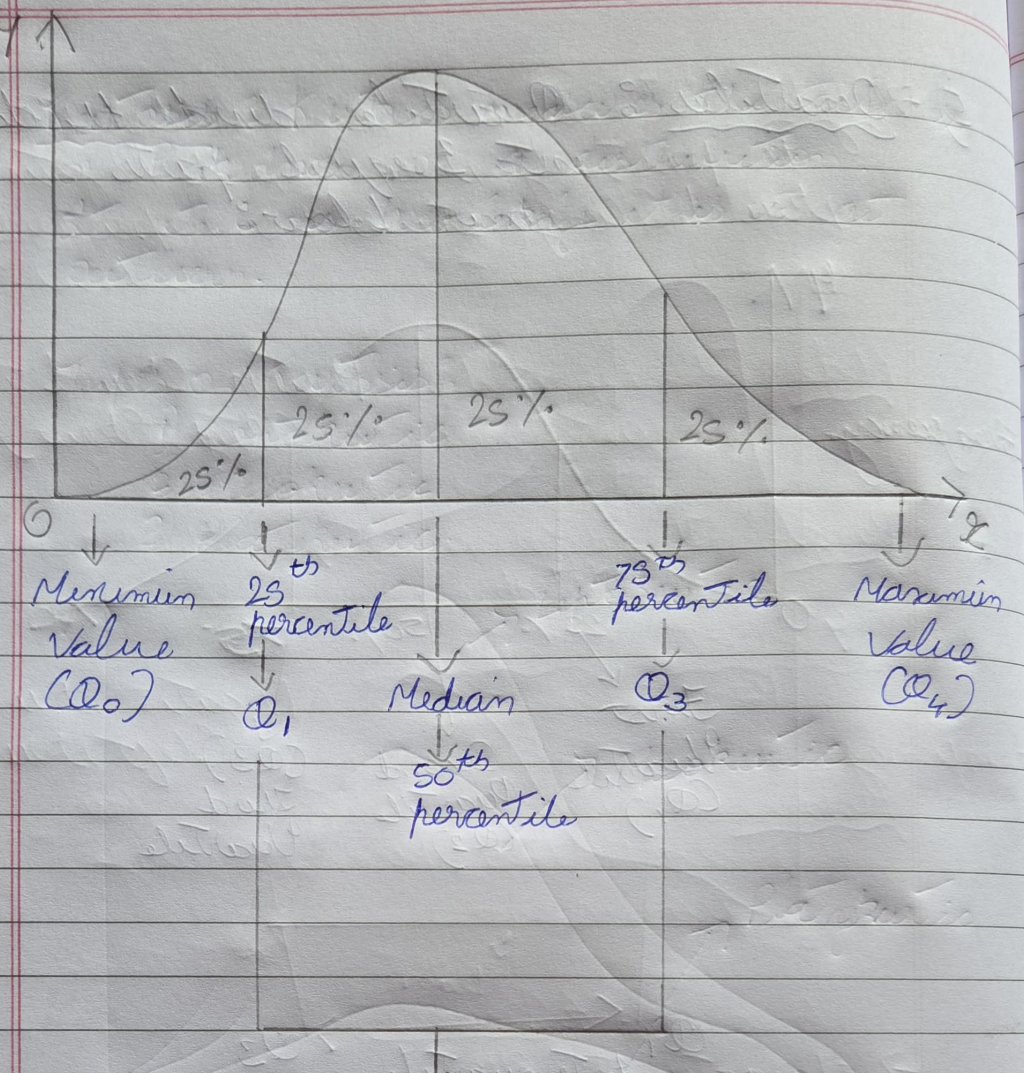


5) Quartiles: Quartiles divide the data into 3 equal parts shown below:



6) Five number Summary: Five number summary gives us the idea of where of the the spread and center of the dataset.

Q_0 = Minimum Value
Q_1 = First Quartile
Q_2 = Median
Q_3 = Third Quartile
Q_4 = Maximum Value.

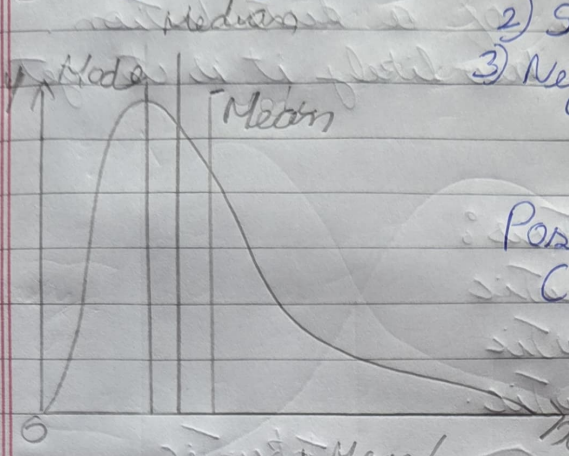


Interquartile Range (IQR)
$= Q_3 - Q_1$

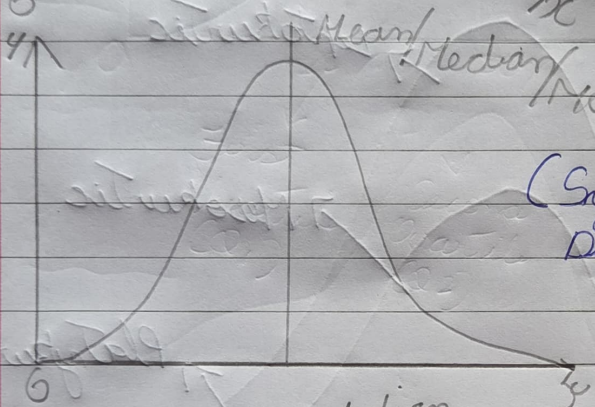
7) Skewness : Skewness is a statistical measure that measures the asymmetry of data distribution.

→ It helps identify if the data is Symmetric or Skewed (Long Tails)

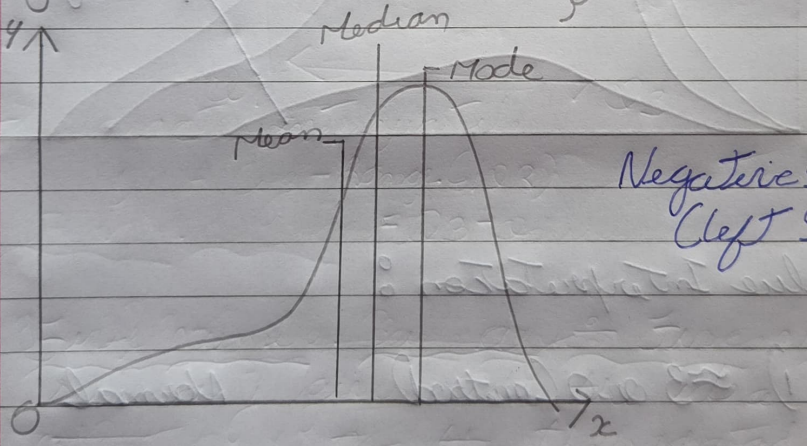
-> Types of Skewness : 1) Positive Skew
2) Symmetrical
3) Negative Skew



Positive Skew
(Right Skew)



(Symmetrical Distribution)



Negative Skew
(left Skew)

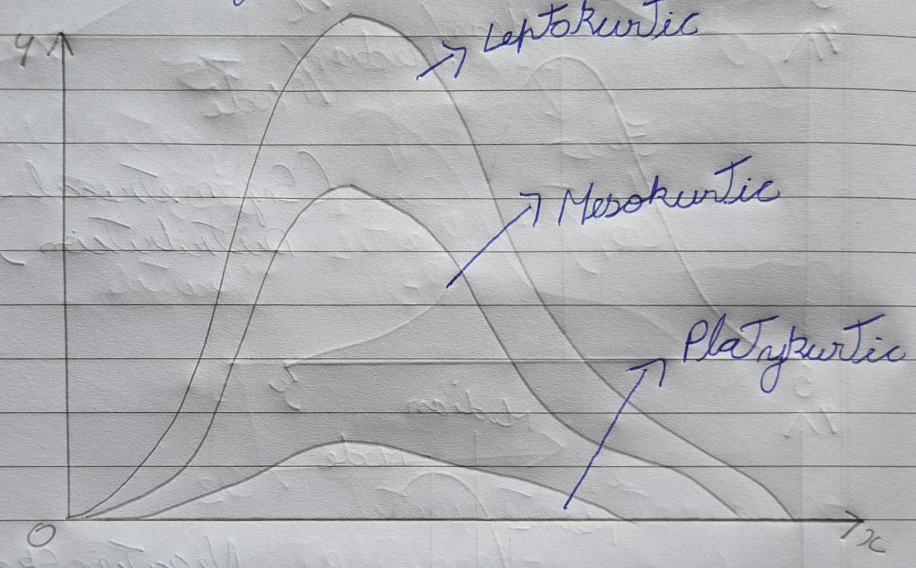
How to calculate Skewness :

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

-> Skewness is important for Model assumptions, Statistical methods and outlier detection.

8) Kurtosis: Kurtosis measures the "tailedness" or "Peakedness" of a distribution. It tell us how likely it is to get outliers.

Types of Kurtosis:
1) Leptokurtic
2) Mesokurtic
3) Platykurtic



Value Interpretation:

1) ≈3 or 0 (Omitted) — Normal
2) >3 or 0 — Leptokurtic
3) <3 or 0 — Platykurtic

Formula to Calculate Kurtosis:

$$\text{Kurtosis} = \frac{E(X-\mu)^4}{\sigma^4}$$

Use Cases:

→ Leptokurtic: Useful to test Models under extreme conditions

→ Platykurtic: Safer but may underestimate risk.

→ Mesokurtic: Ideal baseline.