Probability = How likely something is to happen.

👉 Statistics = How we collect, understand, and learn from data.

0 to 7

3

1

50

w

probability value means 0 to 1

0 head 50 1/2

A random variable is a variable whose value depends on the outcome of a random experiment.

6

$\frac{1}{6}$

1
2
3
4
5
6

Your parYou roll a die → possible outcomes: 1, 2, 3, 4, 5, 6.

Let

X = number that appears on the top.

Here, X is the random variable.

We don’t know the value before rolling → that’s why it’s random.

agraph text

probability distribution

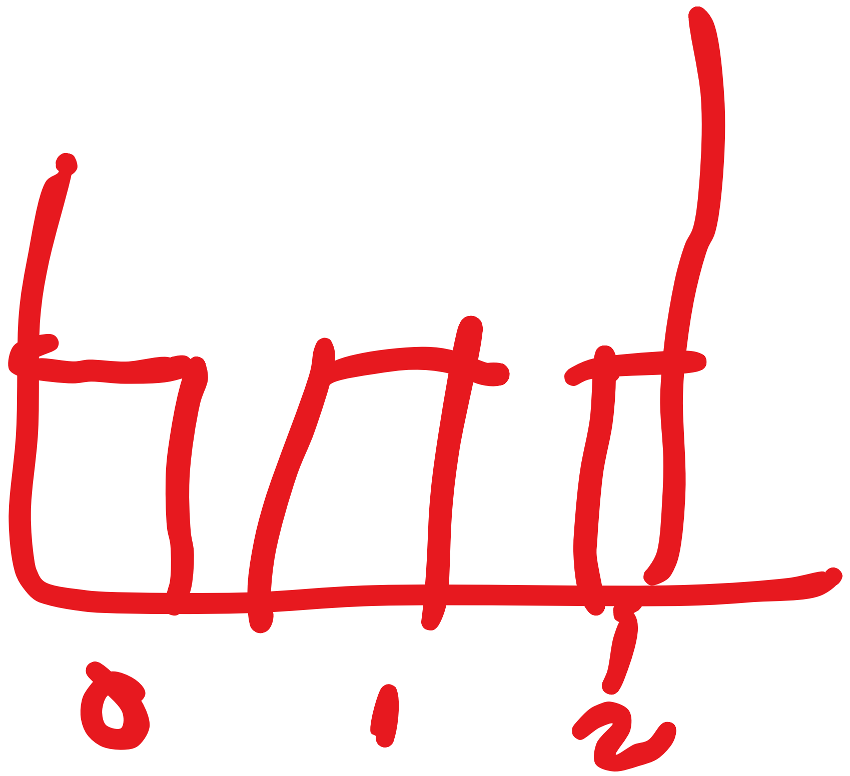
A probability distribution basically tells us how probabilities are spread across all possible outcomes of a random event.

What is Discrete Probability Distribution?

A discrete probability distribution shows all the possible values a random variable can take and how likely each value is.

- “Discrete” → the values are separate, countable, not continuous.
- Probabilities always add up to 1.

Think: like a bar chart of chances.

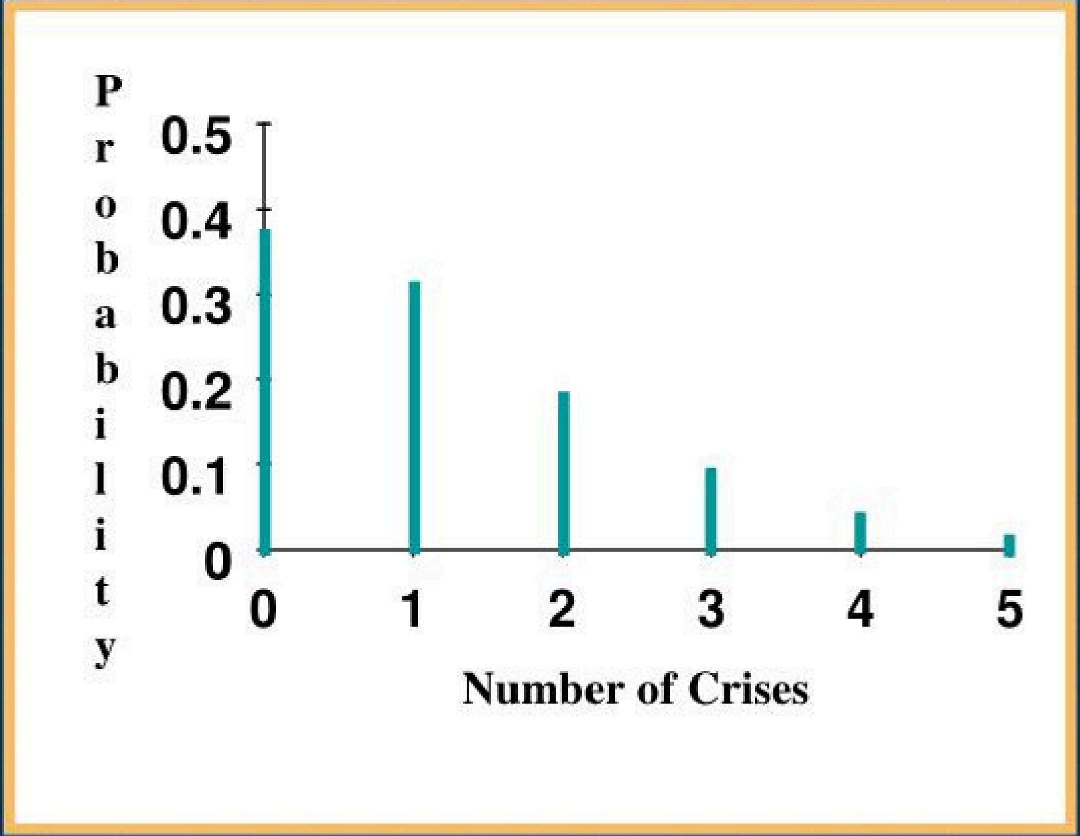


Tossing a coin

- $X = 1$ if Head, 0 if Tail
- $P(X=1) = 0.5$, $P(X=0) = 0.5$

Example: Discrete Distributions & Graphs

Distribution of Daily Crises	
Number of Crises	Probability
0	0.37
1	0.31
2	0.18
3	0.09
4	0.04
5	0.01



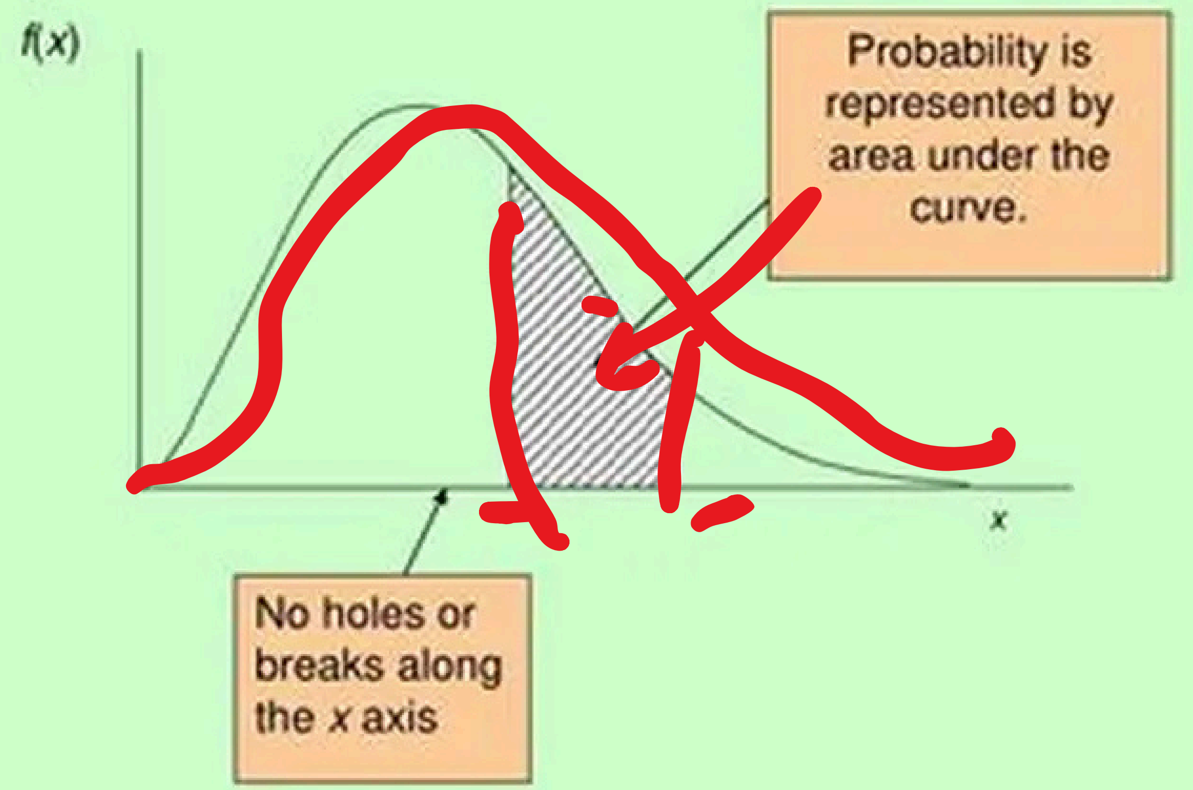

Coca Cola
Coffee
Egg (E-G-G)
Milk
Pizza
Sausages
Sugar
Orange juice
Tea

A continuous probability distribution describes the probabilities of a continuous random variable, which can take any value within a range.

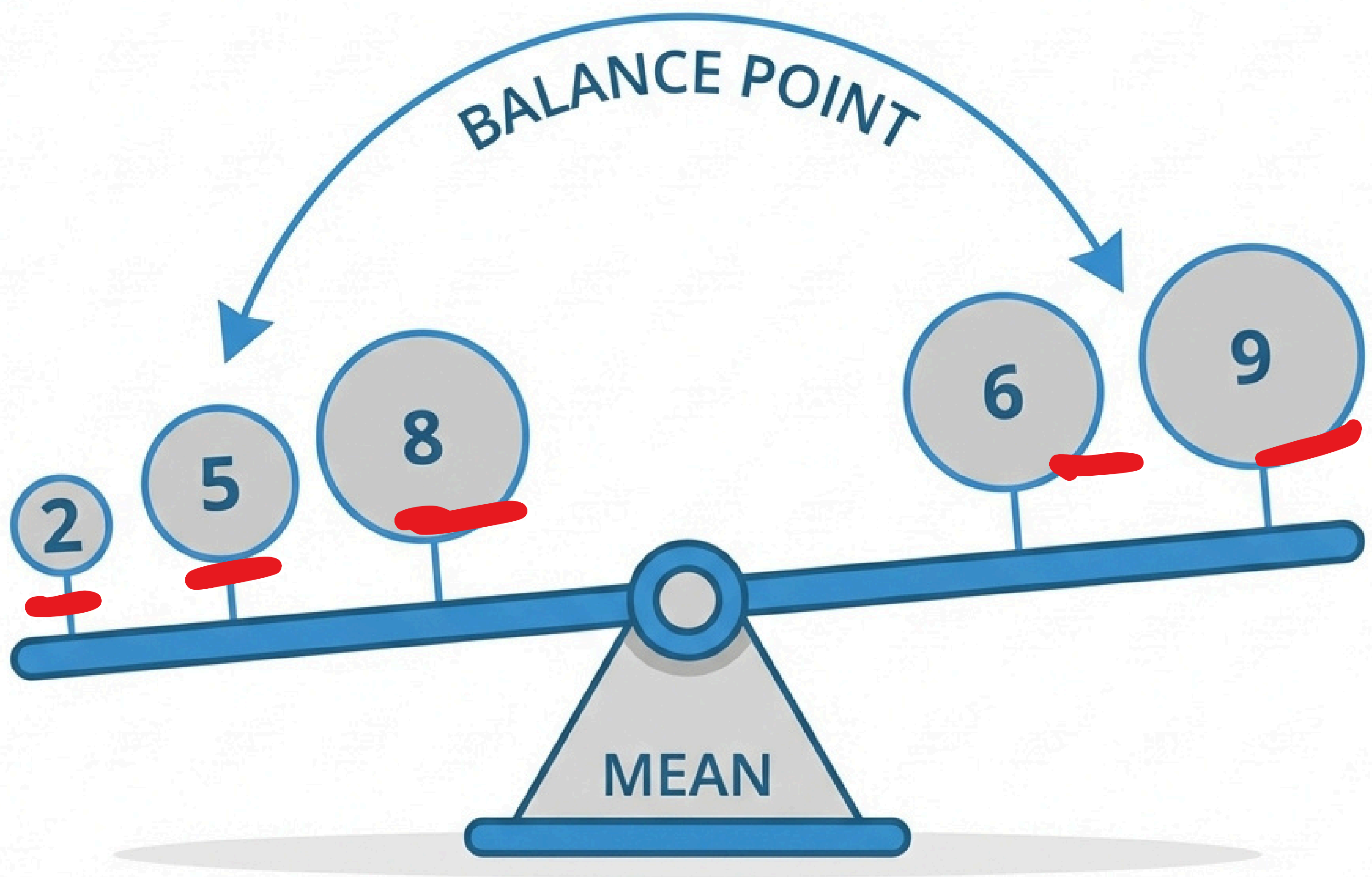
- Unlike discrete variables (countable outcomes like dice), continuous variables are infinite and uncountable.
- Probabilities are measured using areas under a curve instead of exact values.

-10 10 2 6

Figure 6.2 A Continuous Probability Distribution



'Mean'
(Arithmetic Average)

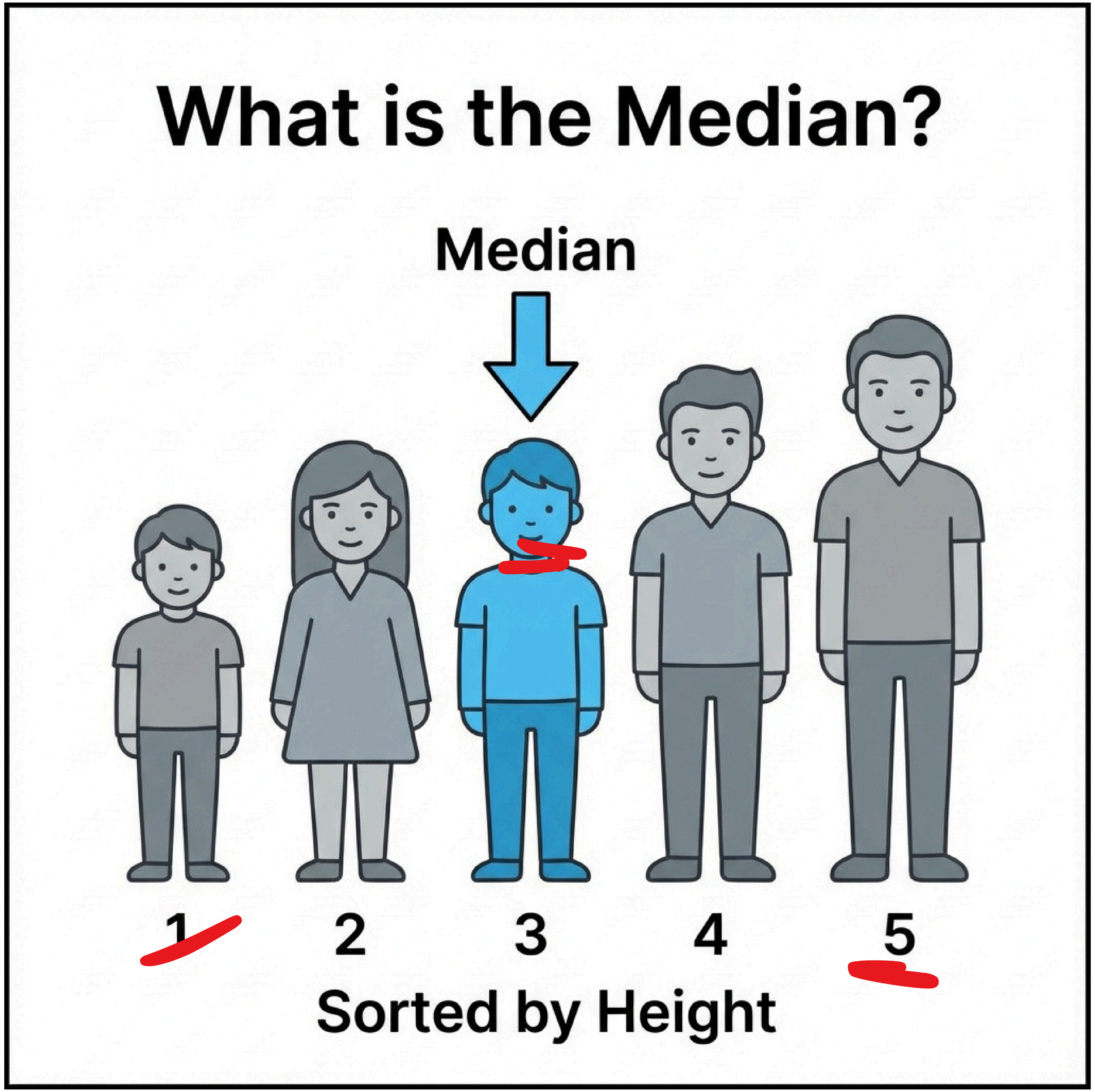


Mean = (2+5+8+6+9) / 5 = 30 / 5 = 6

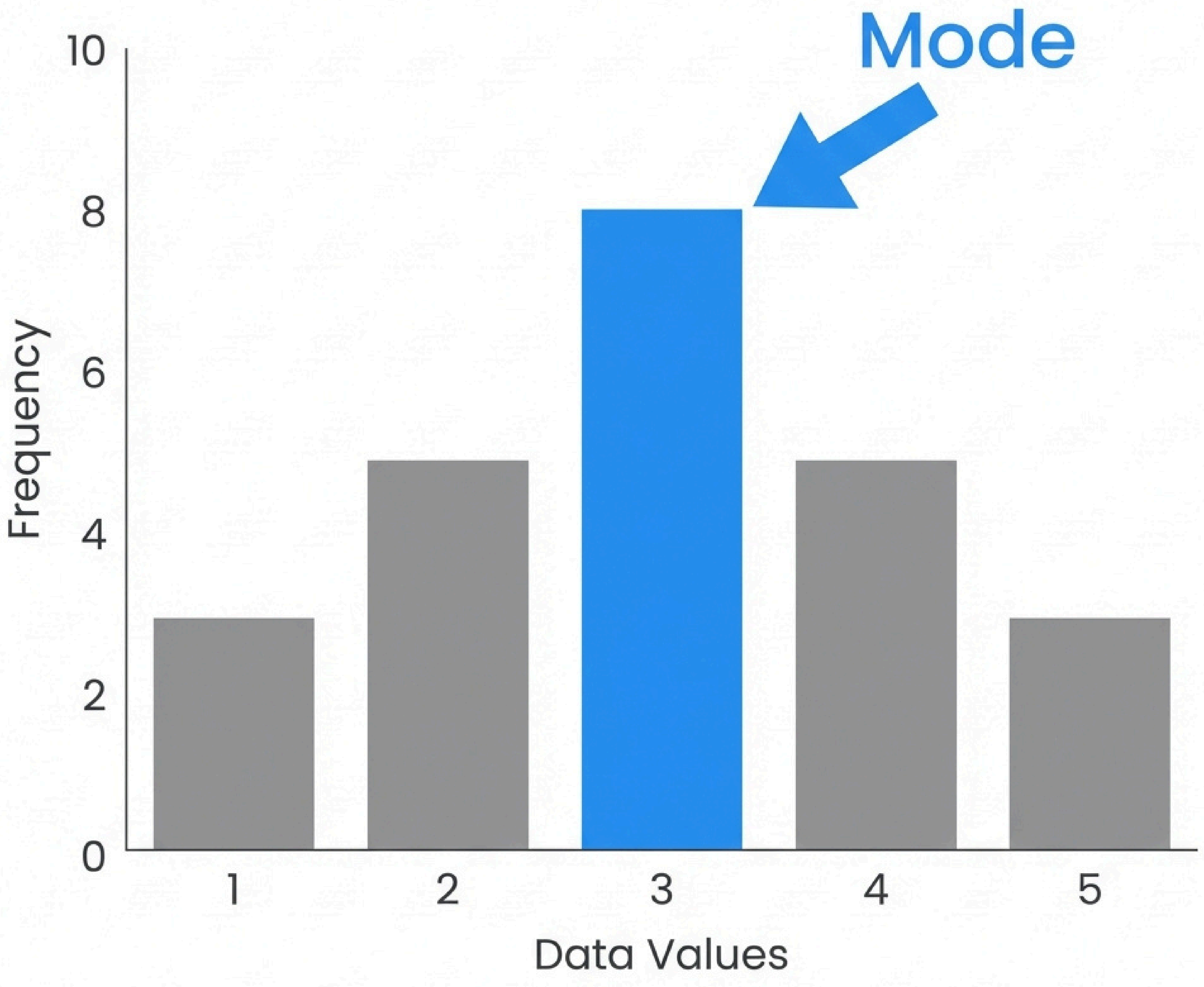
Salary

10k, 20k,
15, 12k
15, 11k

$$\frac{85K}{6} =$$


What is the Median?
Median
1
2
3
4
5
Sorted by Height

Understanding Mode:
The Most Frequent Value



Categorical

Titanic

Embarked

S
C
Q

Row

NaN

Variance

60, 61, 62, 62, 63

62

Data: `[2, 4, 6, 8]`

1. Mean: $\bar{x} = (2+4+6+8)/4 = 5$
2. Deviations: `[2-5, 4-5, 6-5, 8-5] = [-3, -1, 1, 3]`
3. Squared deviations: `[9, 1, 1, 9]`
4. Variance (population):

$$\sigma^2 = \frac{9+1+1+9}{4} = \frac{20}{4} = 5$$

- So variance = 5, meaning the numbers are **spread out from the mean.**

What is Standard Deviation (σ)?

- Standard deviation measures how spread out data is from the mean.
- Smaller σ → data is closer to the mean
- Larger σ → data is more spread out

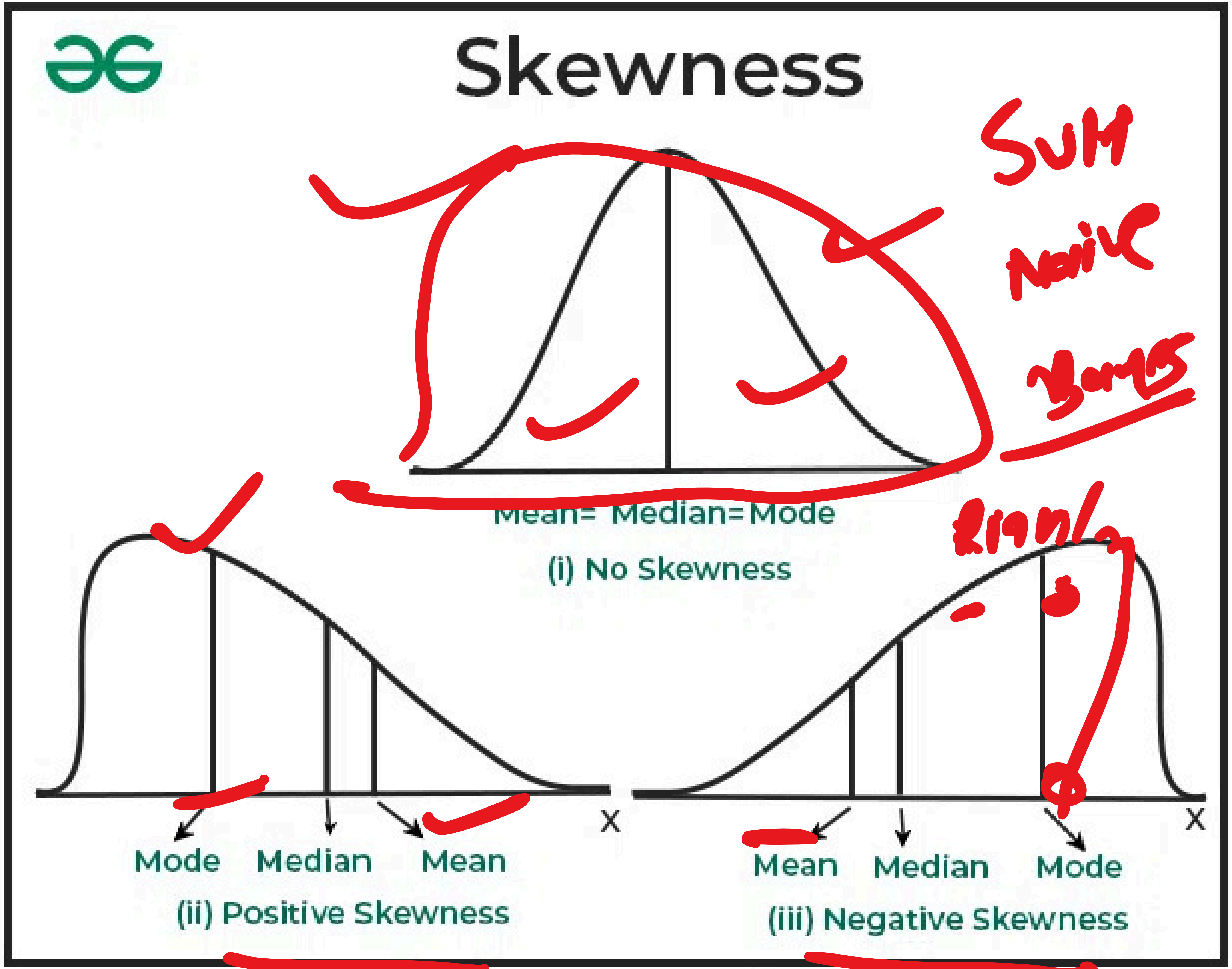
Think of it as:

"How tightly the data points hug the mean."

60, 61, 61

What is Skewness?

- Skewness measures the asymmetry of a data distribution.
- It tells us whether the data is tilted to the left, right, or balanced.
- Basically:
  - Is the data symmetrical?
  - If not, which side has the "long tail"?

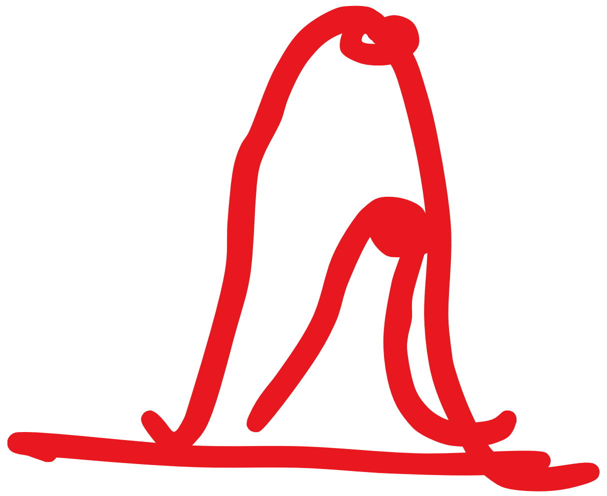

Skewness
Mean= Median=Mode
(i) No Skewness
X
Mode
Median
Mean
(ii) Positive Skewness
X
Mean
Median
Mode
(iii) Negative Skewness

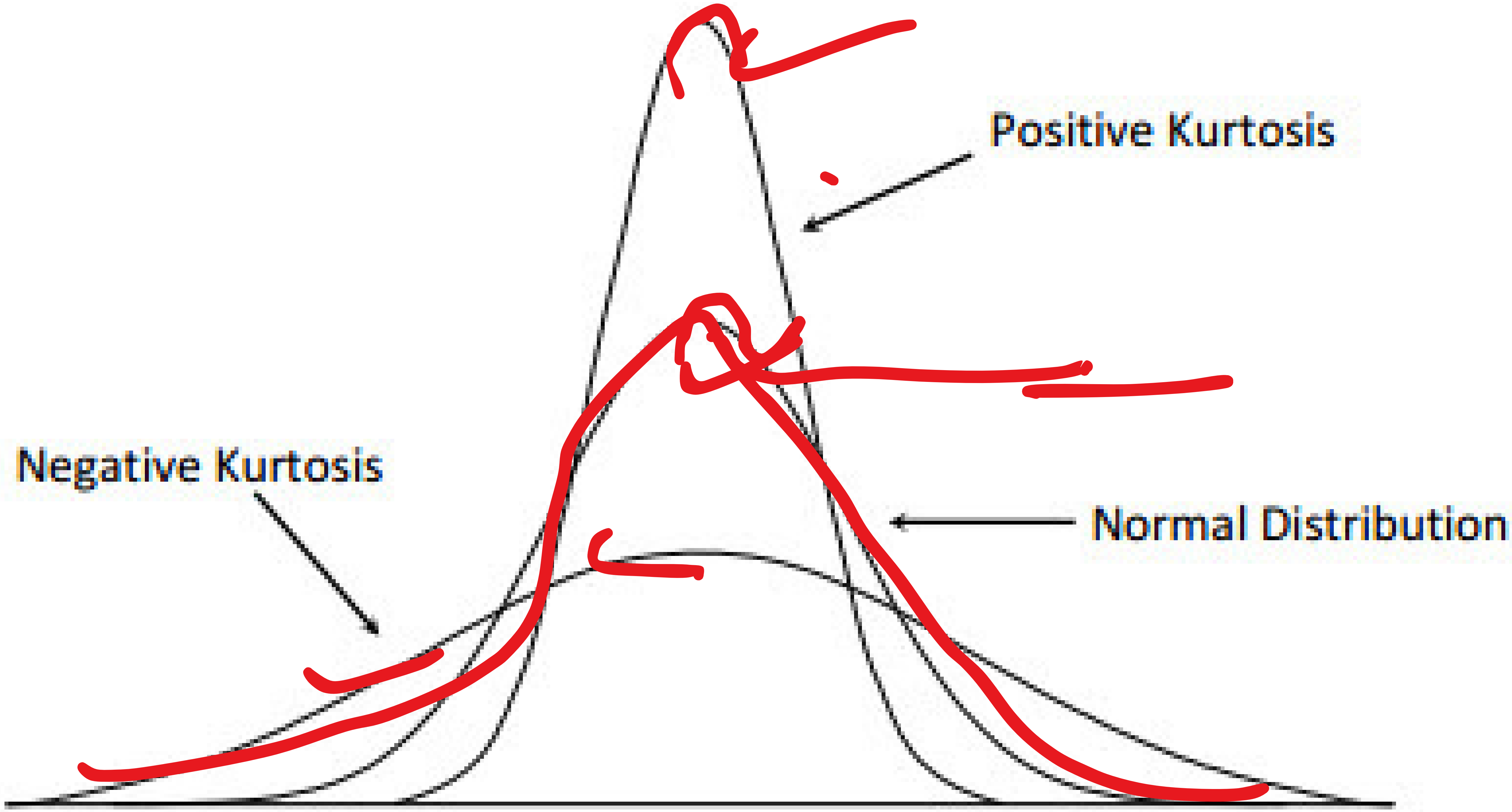
What is Kurtosis?

- Kurtosis measures the “peakedness” or “flatness” of a probability distribution.
- It tells us how heavy or light the tails of the distribution are compared to a normal distribution.

In simple words:

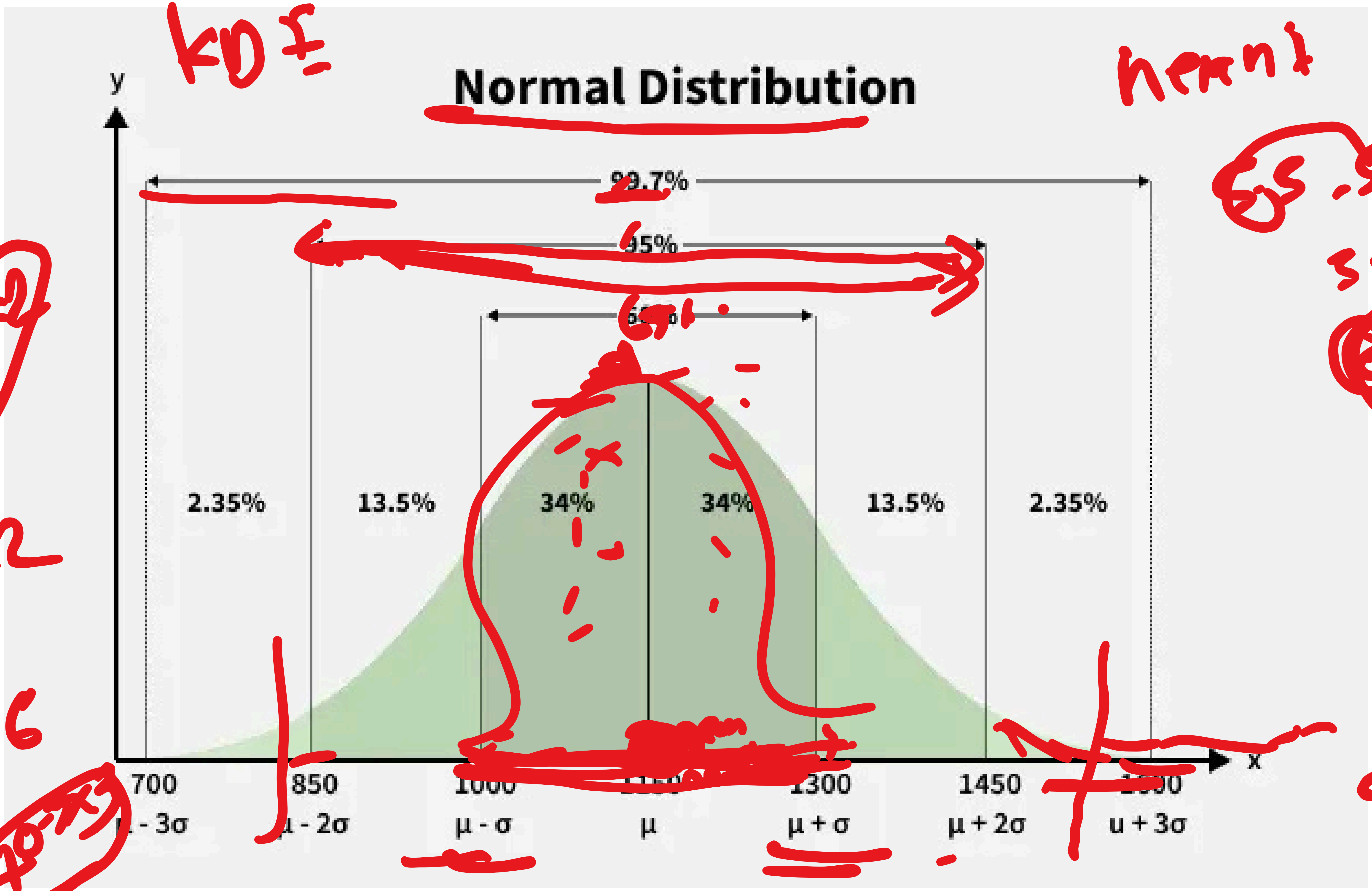
It shows whether the data has more extreme values (outliers) or is flatter.





Positive Kurtosis
Negative Kurtosis
Normal Distribution

- Normal distribution → kurtosis ≈ 3 (mesokurtic)
- Dataset with many extreme high/low values → kurtosis > 3 (leptokurtic)
- Dataset with evenly spread values → kurtosis < 3 (platykurtic)


Normal Distribution
y
x
99.7%
95%
68%
2.35%
13.5%
34%
34%
13.5%
2.35%
700
μ - 3σ
850
μ - 2σ
1000
μ - σ
μ
1300
μ + σ
1450
μ + 2σ
u + 3σ

- Many natural phenomena follow it (heights, weights, errors).
- Symmetric → mean = median = mode → easy analysis.
- Basis of statistical tests (t-test, ANOVA, regression).
- Predictable (68–95–99.7 rule) → understand spread and outliers.
- Central Limit Theorem → sums/averages of variables → normal.
- Widely used in ML → Gaussian Naive Bayes, probabilistic models, density estimation.

Q1 = 33

1️⃣ 25th Percentile (Q1)

Q1

- Also called **first quartile (Q1).**
- **25% of the data is less than this value**, 75% is more.
- Think of it as the "**lower boundary of the lower quarter**".

Example:

Data (sorted): 10, 12, 15, 18, 20, 22, 25, 30

Q1 = 19

- **25th percentile ≈ 14 → 25% of numbers are ≤ 14**

20, 21, 22, 23, 24, 23, 22,

32

2️⃣ 75th Percentile (Q3)

- Also called **third quartile** (Q3).
- 75% of the data is **less** than this value, 25% is more.
- Think of it as the **"upper boundary of the upper quarter"**.

Q3

Example (same data):

- 75th percentile ≈ 24 → **75% of numbers are ≤ 24**

upper

Q1 Q3 75%.

Q3 = 24

lower

Q1 = 14

Q1 Q3

Q3 = 24,

2️⃣ Using IQR (Interquartile Range) – Very Popular

- Step 1: Find Q1 (25th percentile) and Q3 (75th percentile)
- Step 2: Calculate IQR:

$$IQR = Q3 - Q1$$

- Step 3: Define outlier limits:

$$\text{Lower limit} = Q1 - 1.5 \times IQR$$

$$\text{Upper limit} = Q3 + 1.5 \times IQR$$

- Data outside these limits = **outliers**

Role

Age

20

30

40

28

27

0 - 1

Fare

Train

50

60

70

300

4000

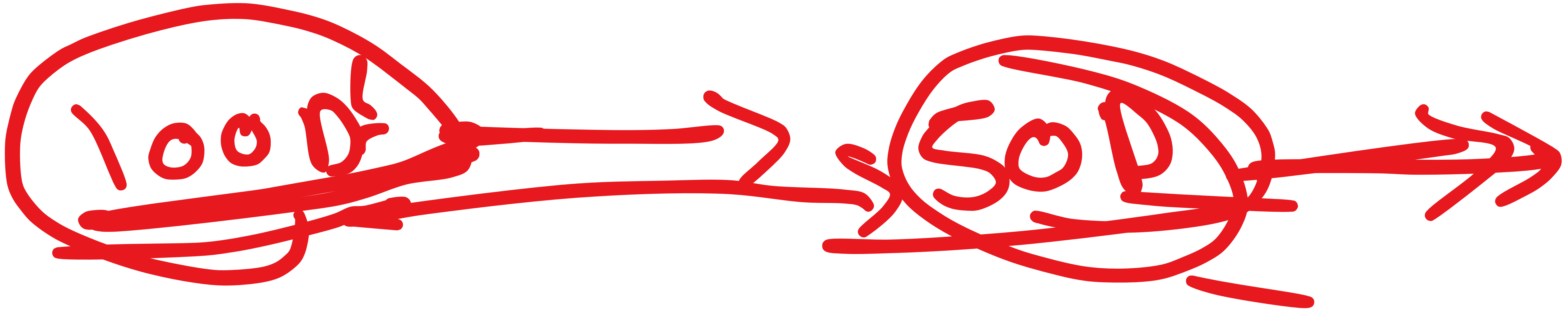
5000

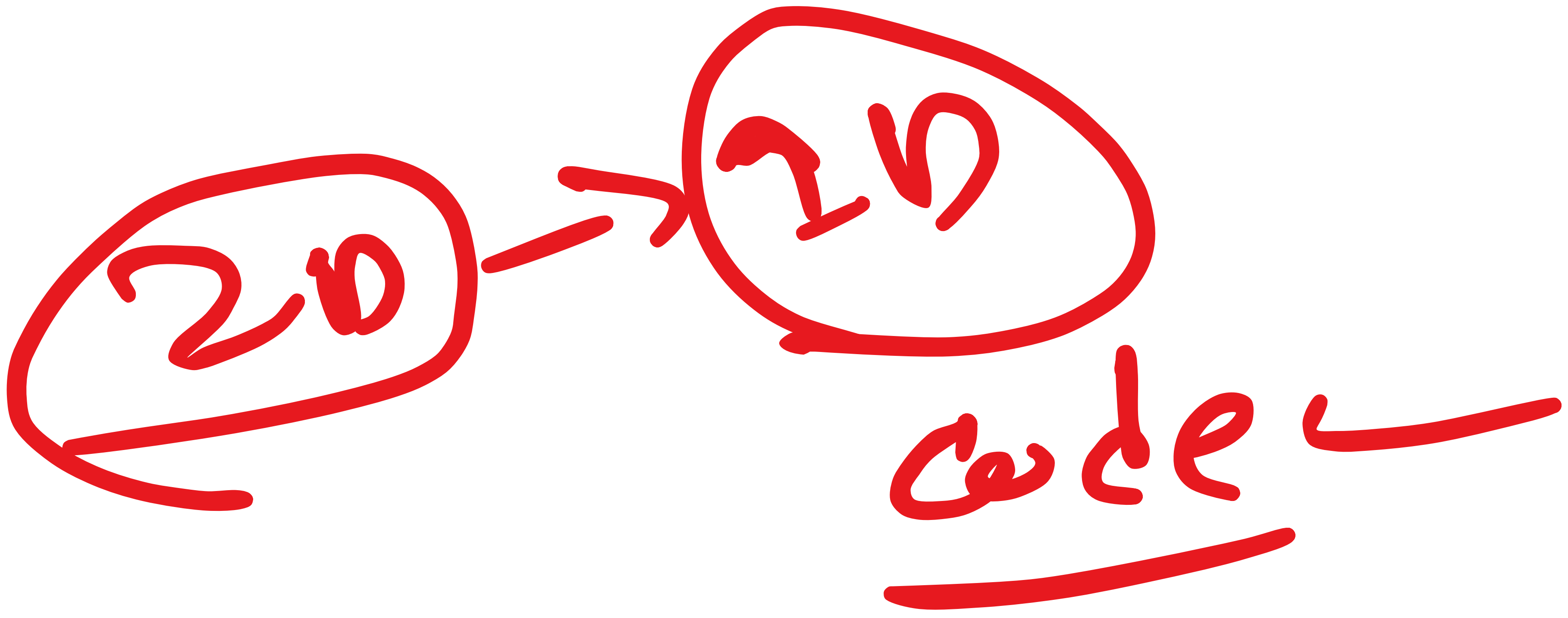
0 - 7

weight

Height

BMI


1000
500


20
20
code

Data

Sign language

70000

Hospital

open cv

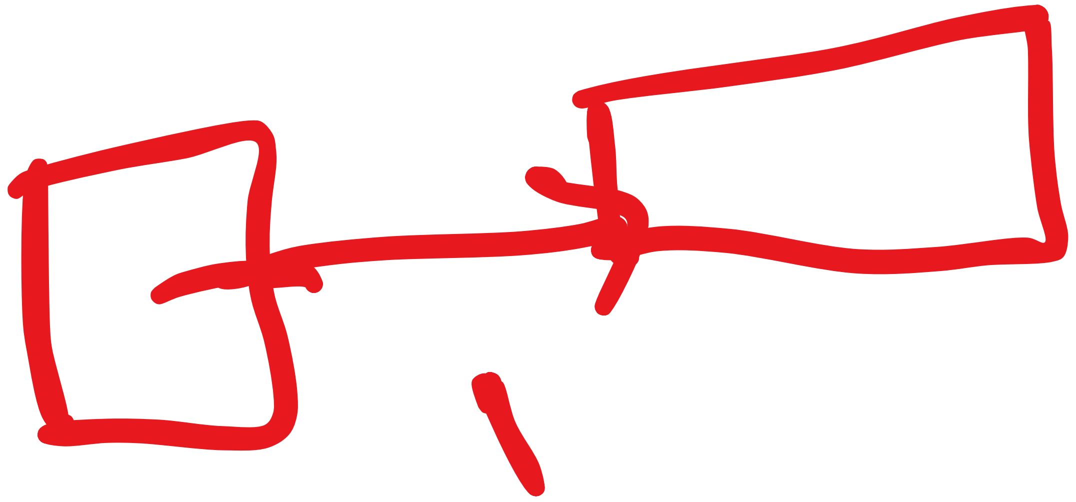
400-500

Kerlo

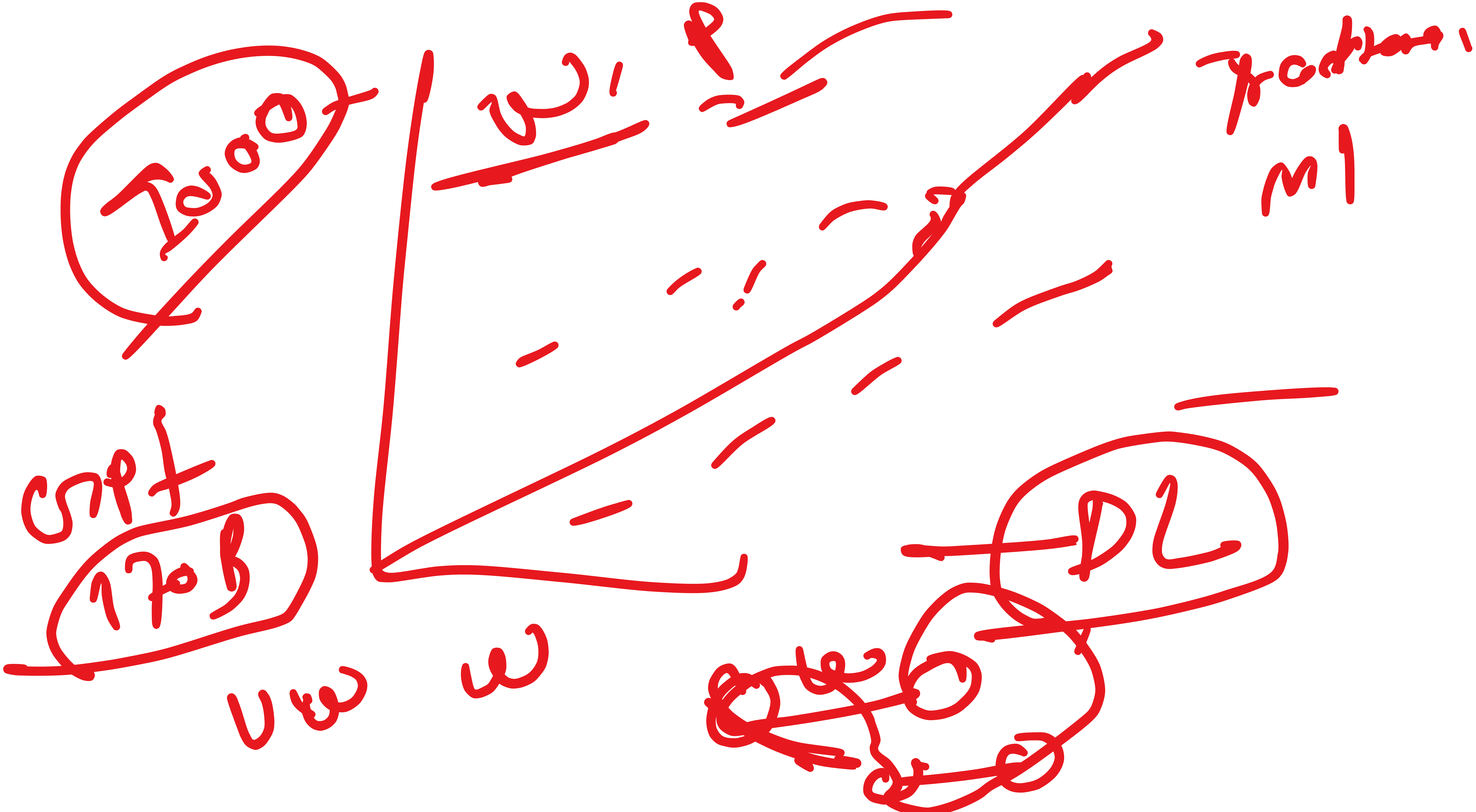
sign

man

Normale



30%
2
2.5
28
54
53
27
0
58
56
7
60, 61
82
8
34
22



Normalization (Important)

Data

ml model

Fare (dense)

0-1

0-1

Age

20

30

25

40

Fare

300

100

400

500

7)

MinmaxScaler()

$$\text{Normalization} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad 0-1$$

9, 10, 11, 12, 13

$$\frac{10 - 9}{13 - 9} = \frac{1}{4}$$