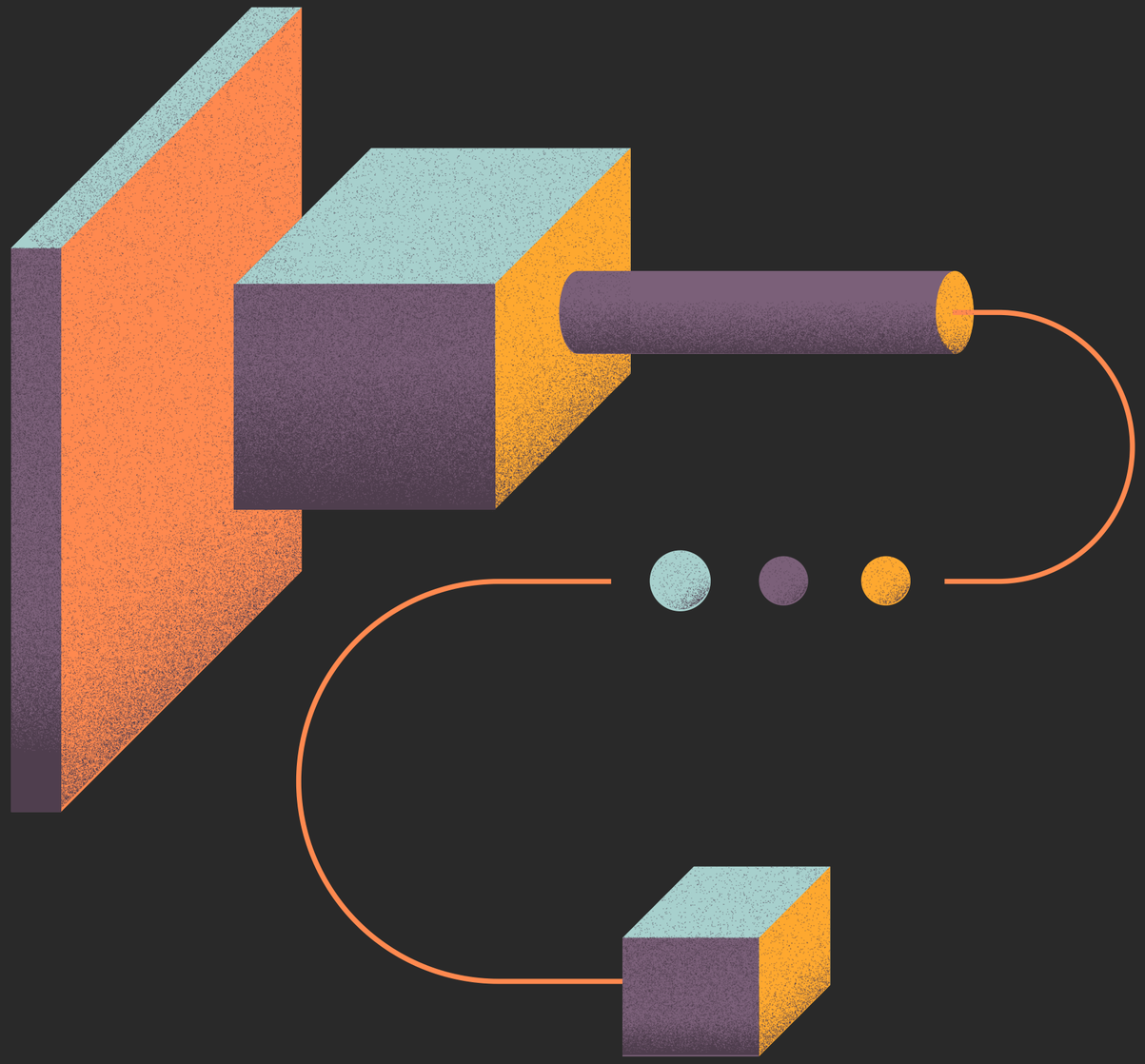
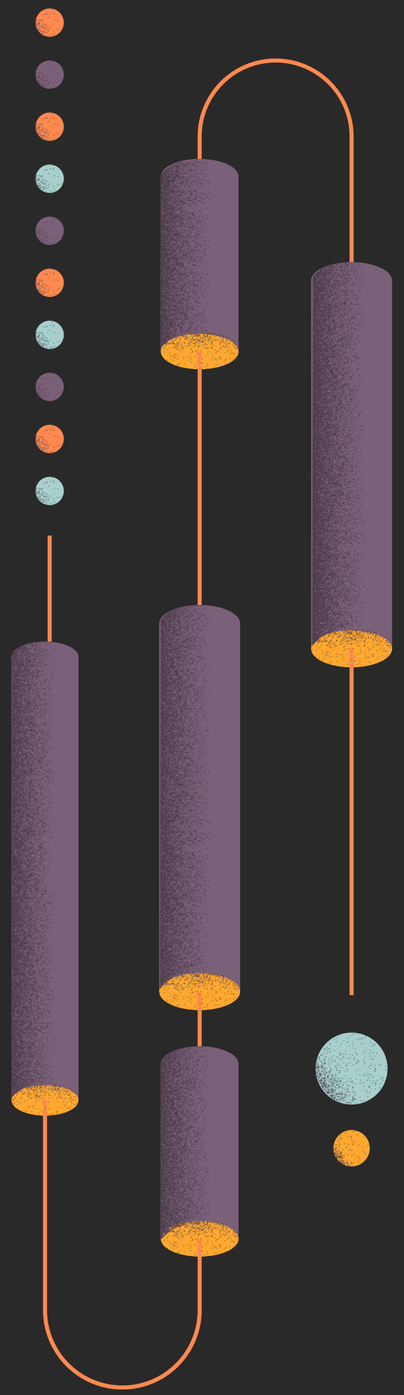
DATA ANALYSIS + STATISTICS



Data Science Club



VOCABULARY

Data Science Statistics

- Population vs Sample
- Measures of central tendency: Mean, Mode, Median
- Standard deviation
- Z-Score
- Correlation coefficient
- Correlation vs Causation

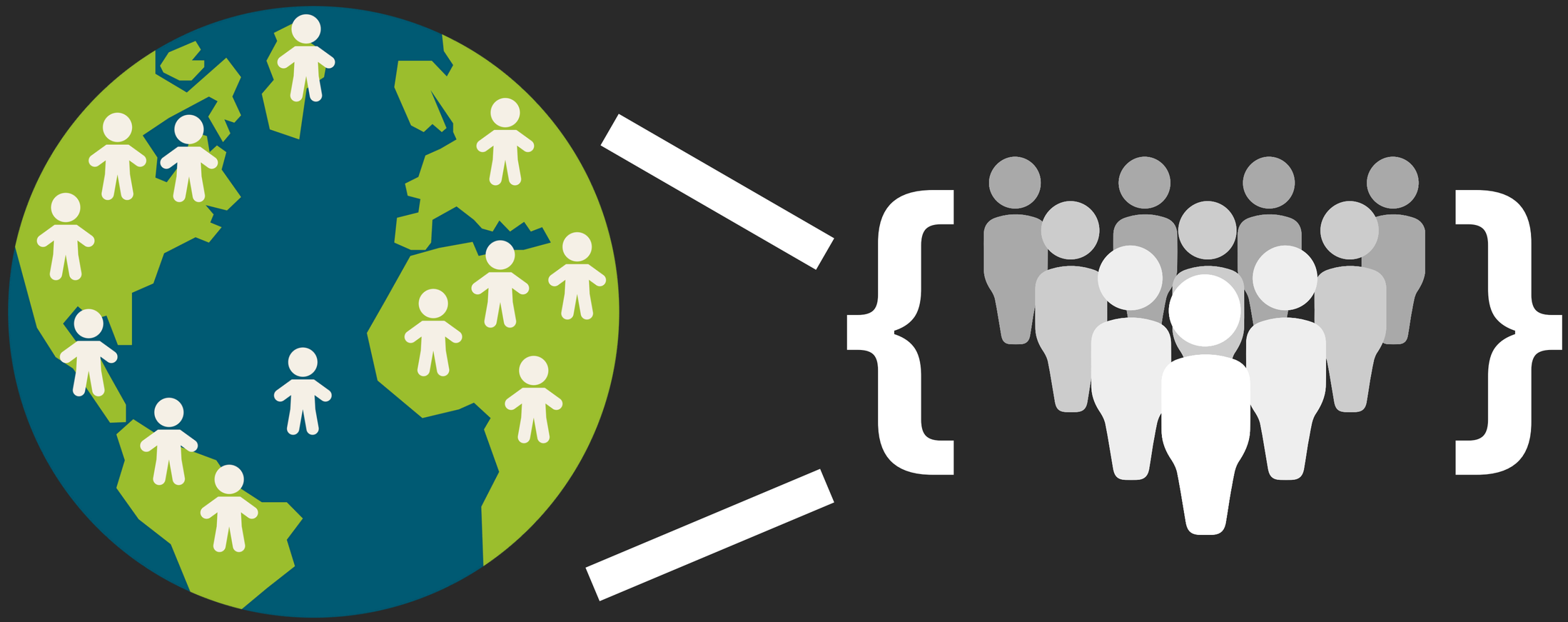
2 types of data you can collect for analysis

Statistical

- Finding trends and patterns
- Numerical data for graphing, organizing and drawing conclusions
- Quantitative

Non-statistical

- Provides generic information
- Sound, images, texts, etc.
- Qualitative



POPULATION VS. SAMPLE

MEASURES OF CENTRAL TENDENCY:

MEAN

Another name for average

E.g. What is the mean of 2,5,7,8?

Mean: (2+5+7+8)/4 = 5.5

MEDIAN

The middle number

Even: Average of two numbers

Odd: Pick middle number

MODE

The number that appears the most

E.g. 7,3,5,7,7,8,7,7,9

Mode: 7

MEASURES OF CENTRAL TENDENCY:

What is the mean?

1,52,65,2,1002

MEASURES OF CENTRAL TENDENCY:

What is the mean?

1,52,65,2,1002

Answer: 224.4

MEASURES OF CENTRAL TENDENCY:

What is the mode?

2,4,5,1,2,3,2,3,3

MEASURES OF CENTRAL TENDENCY:

What is the mode?

2,4,5,1,2,3,2,3,3

Answer: 2 and 3
It is a BIMODAL data.

MEASURES OF CENTRAL TENDENCY:

What is the median?

1,2,3,4,5,6,7,8

MEASURES OF CENTRAL TENDENCY:

What is the median?

1,2,3,4,5,6,7,8

Answer: 4.5
Take the average of
4 and 5.

Data set 1

{2,3,4,5,6}

Data set 2

{-10,-1,4,9,18}

Find the mean

Data set 1

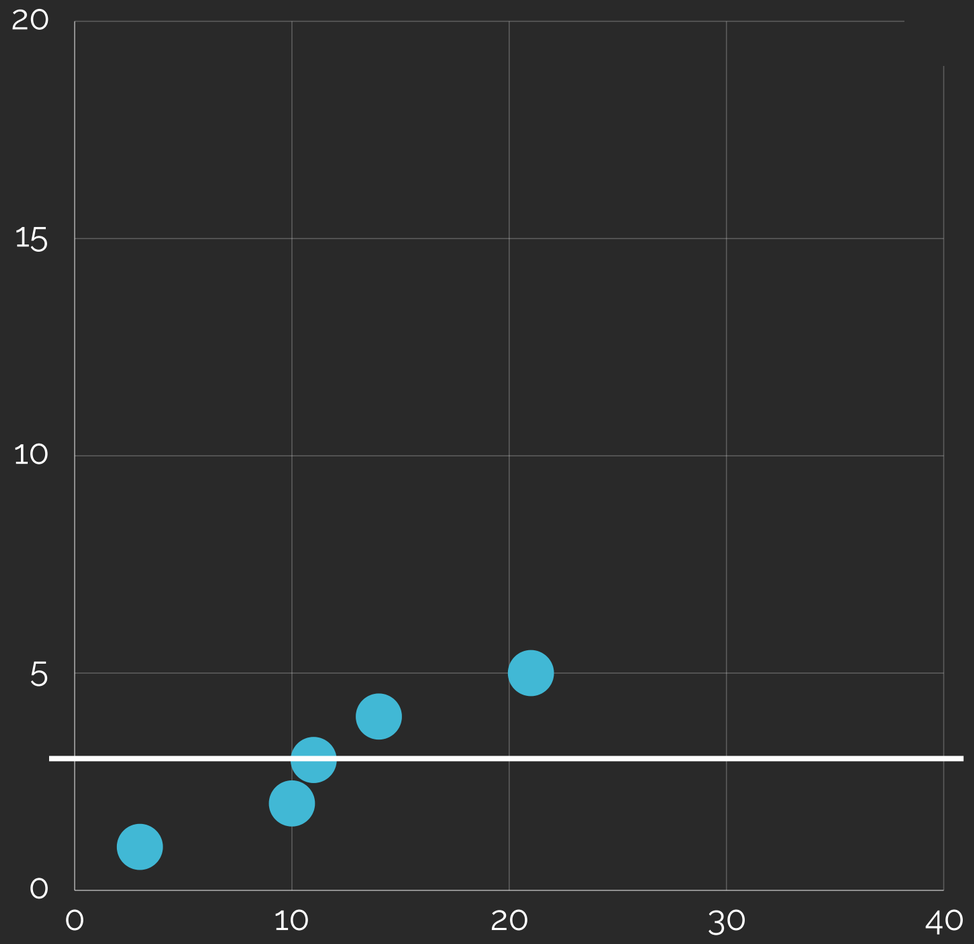
{2,3,4,5,6}

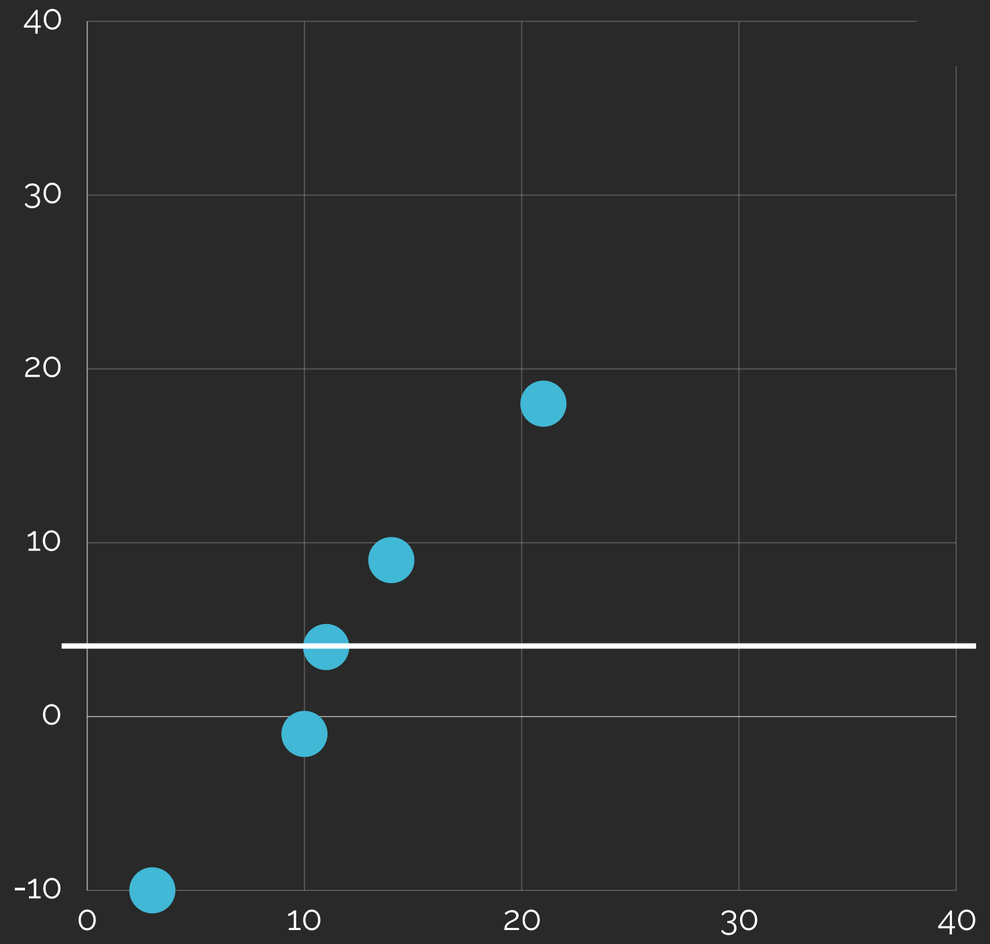
Mean = 4

Data set 2

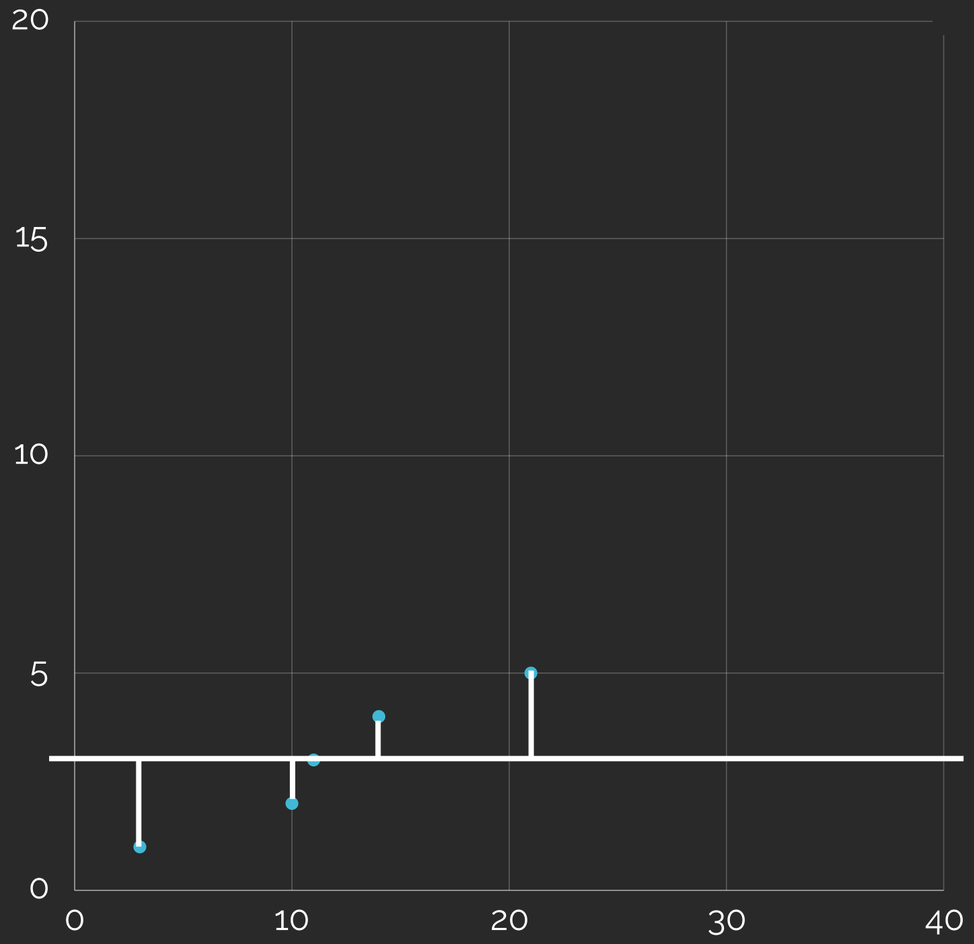
{-10,-1,4,9,18}

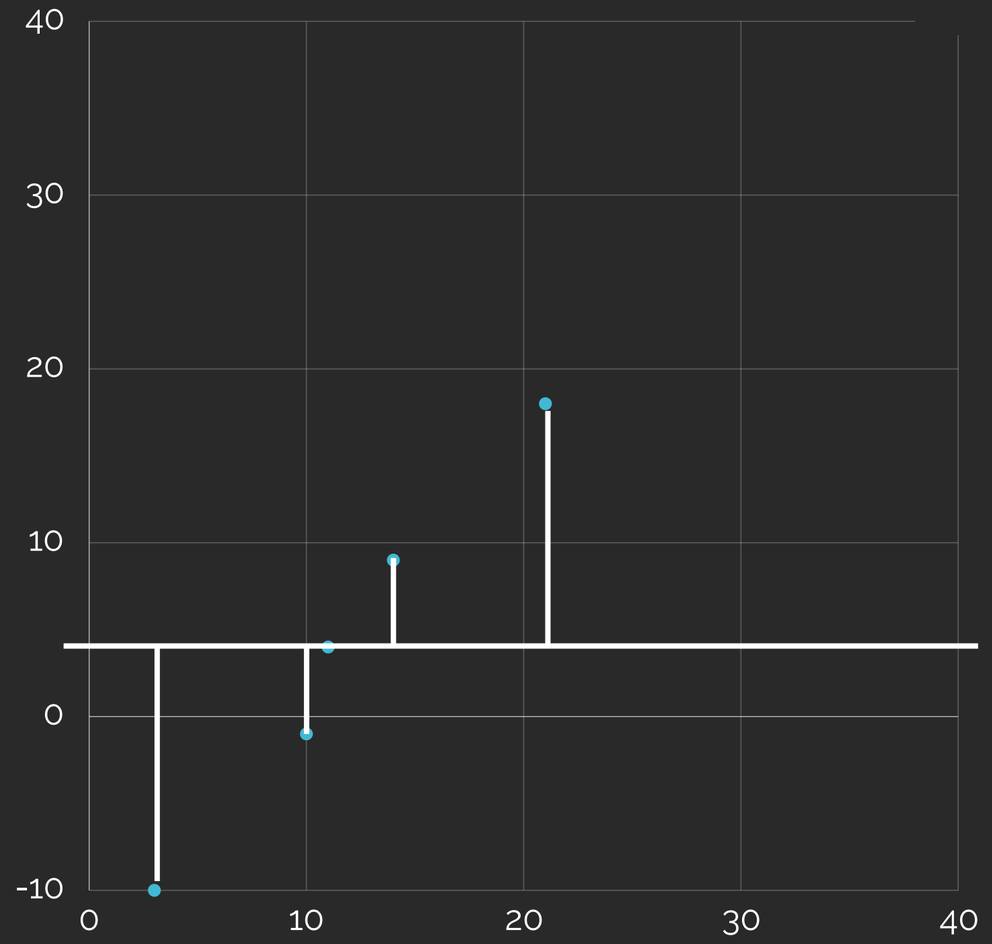
Mean = 4


20
15
10
5
0
0
10
20
30
40


40
30
20
10
0
-10
0
10
20
30
40

How far each observation is from the mean





Deviations from mean

$$x_i - \bar{x}$$

Observation

Mean

For population

Observation

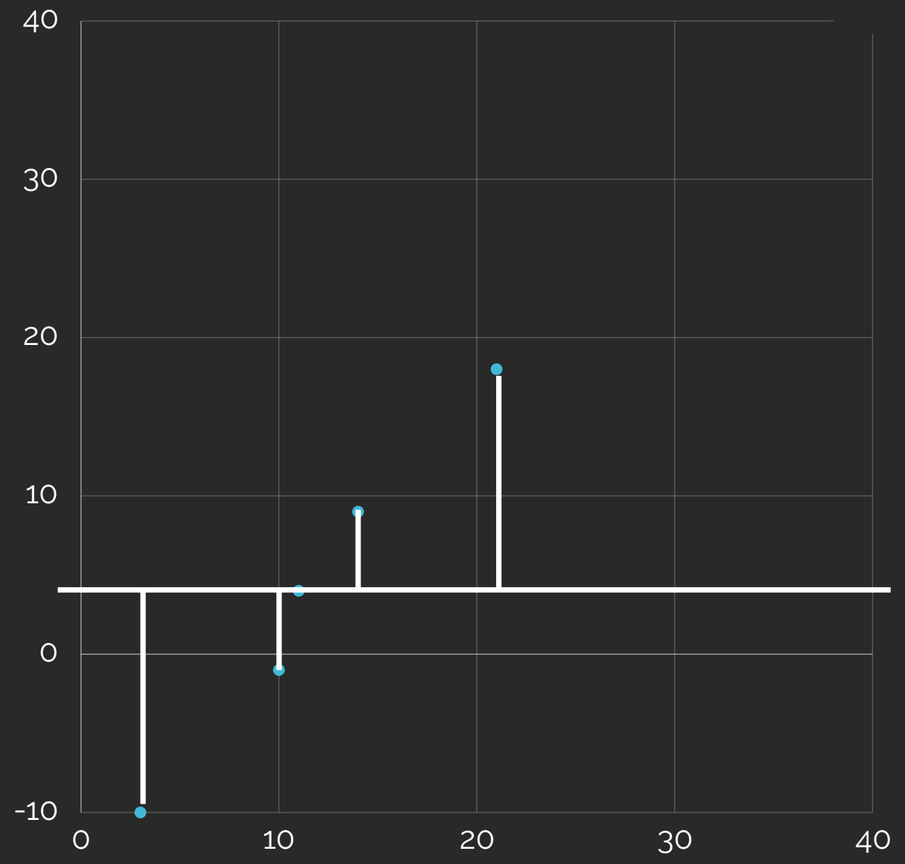
Population mean (“Mu”)

$$x_i - \mu$$

If all distance in magnitude are small (close), less variability

If all distance in magnitude are large (far away), high variability

Average of deviation must always be 0 (since sum will be 0)



Square it!

$$(x_i - \mu)^2$$

- Always gives non-negative values
- Demonstrated the strength by emphasizing the difference

Divide by size of population

$$\frac{(X_i - \mu)^2}{N}$$

- Use n-1 for sample to reduce bias

STANDARD DEVIATION

Denoted by “σ”, pronounced as “sigma”

Square root it

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ = population standard deviation

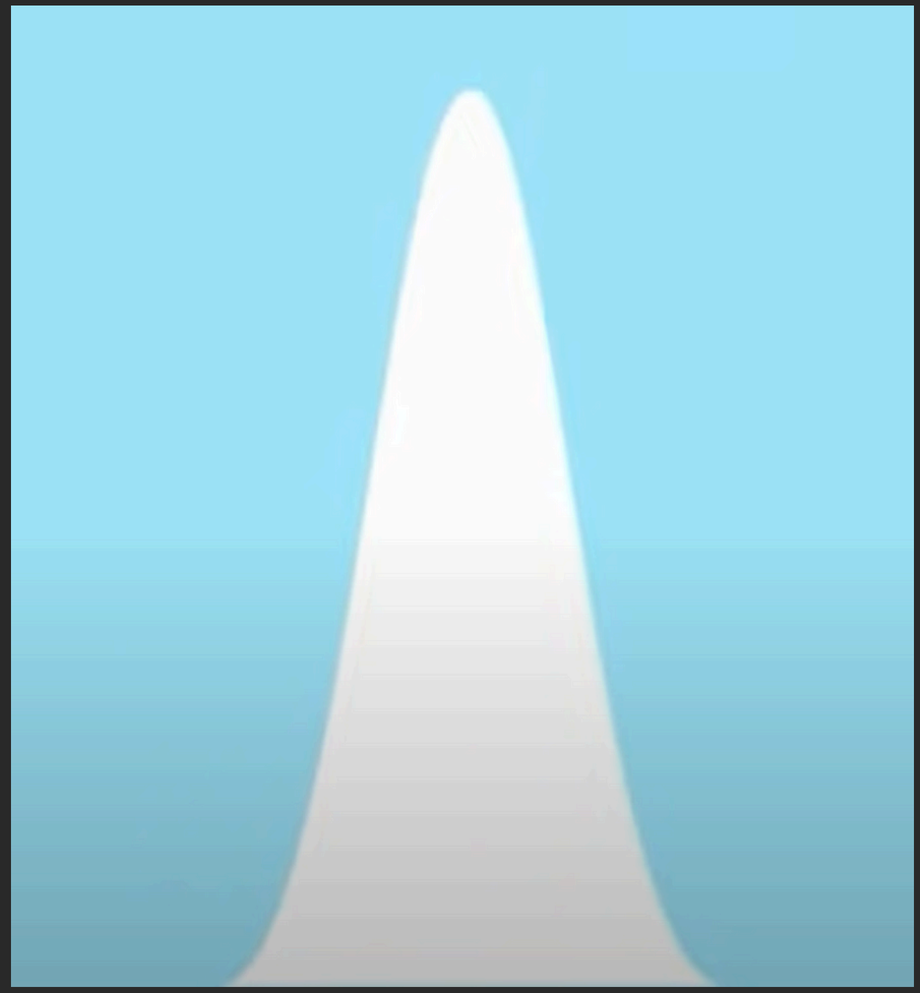
N = the size of the population

x_i = each value from the population

μ = the population mean

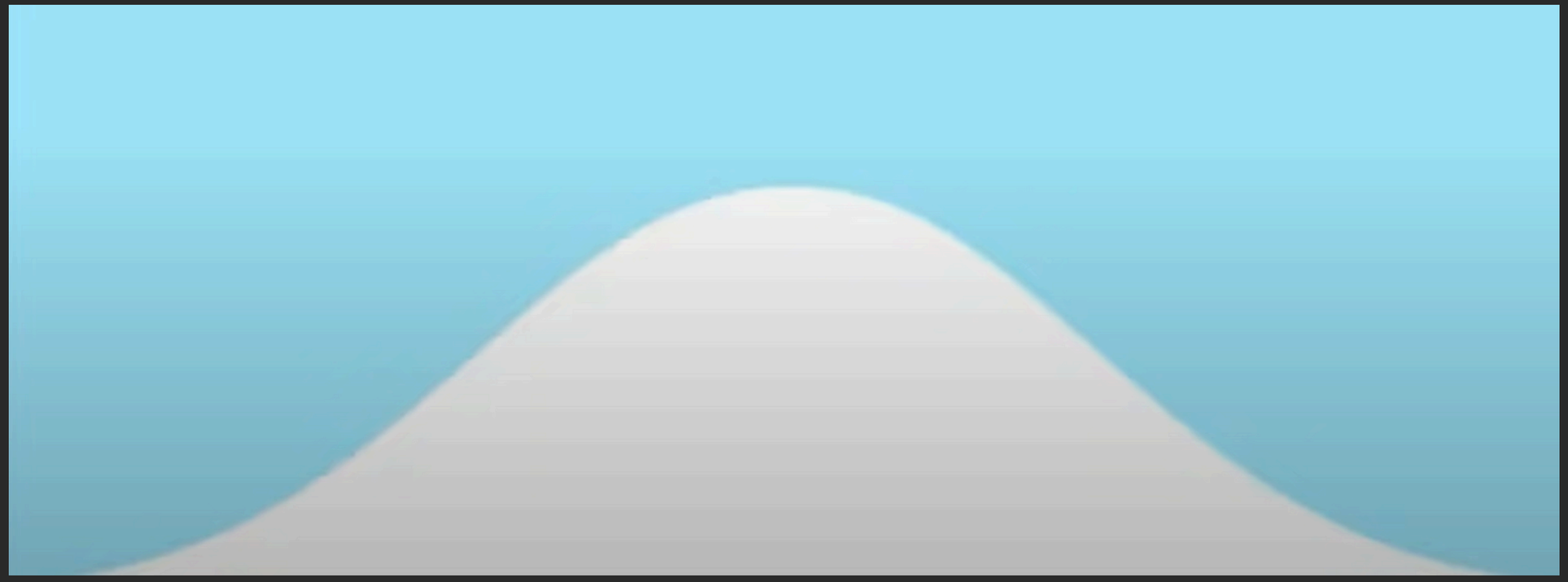
LOW STANDARD DEVIATION

CLUSTERED AROUND THE MEAN



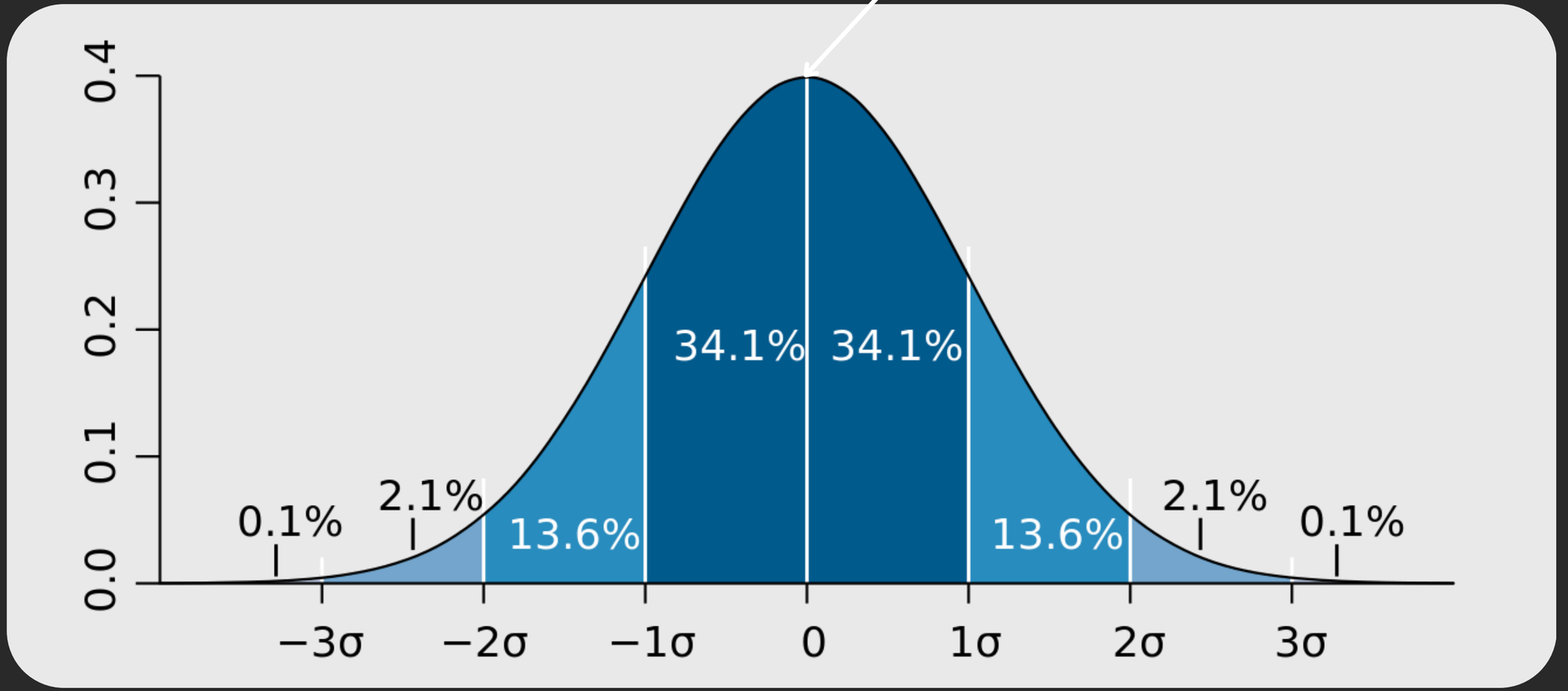
HIGH STANDARD DEVIATION

SPREAD AWAY FROM THE MEAN



NORMAL DISTRIBUTION

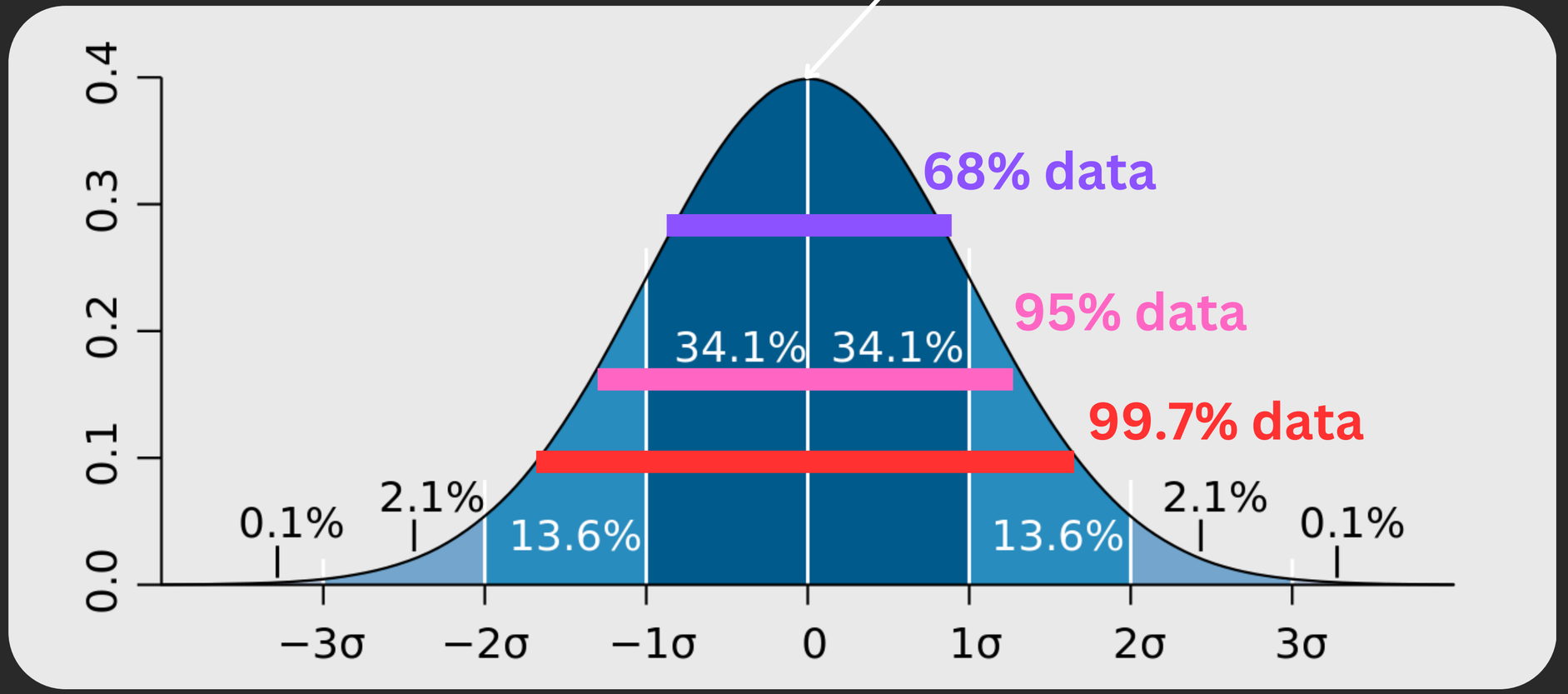
BELL CURVE/ GAUSSIAN DISTRIBUTION



Significant values only 3 SD below/above mean

NORMAL DISTRIBUTION

BELL CURVE/ GAUSSIAN DISTRIBUTION

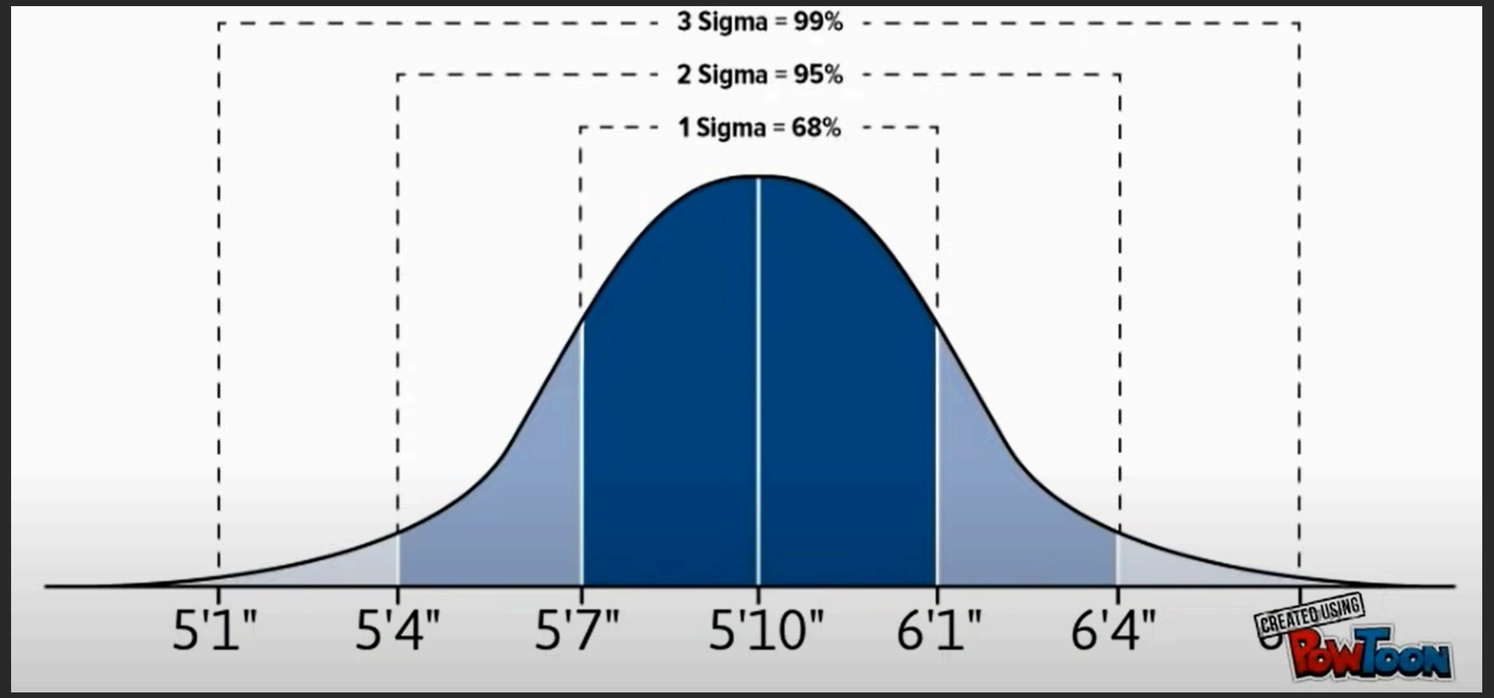


Significant values only 3 SD below/above mean

NORMAL DISTRIBUTION

ALLOWS US TO USE PROBABILITY TO UNDERSTAND IF A VALUE IS EXPECTED TO NEEDS FURTHER INVESTIGATION.

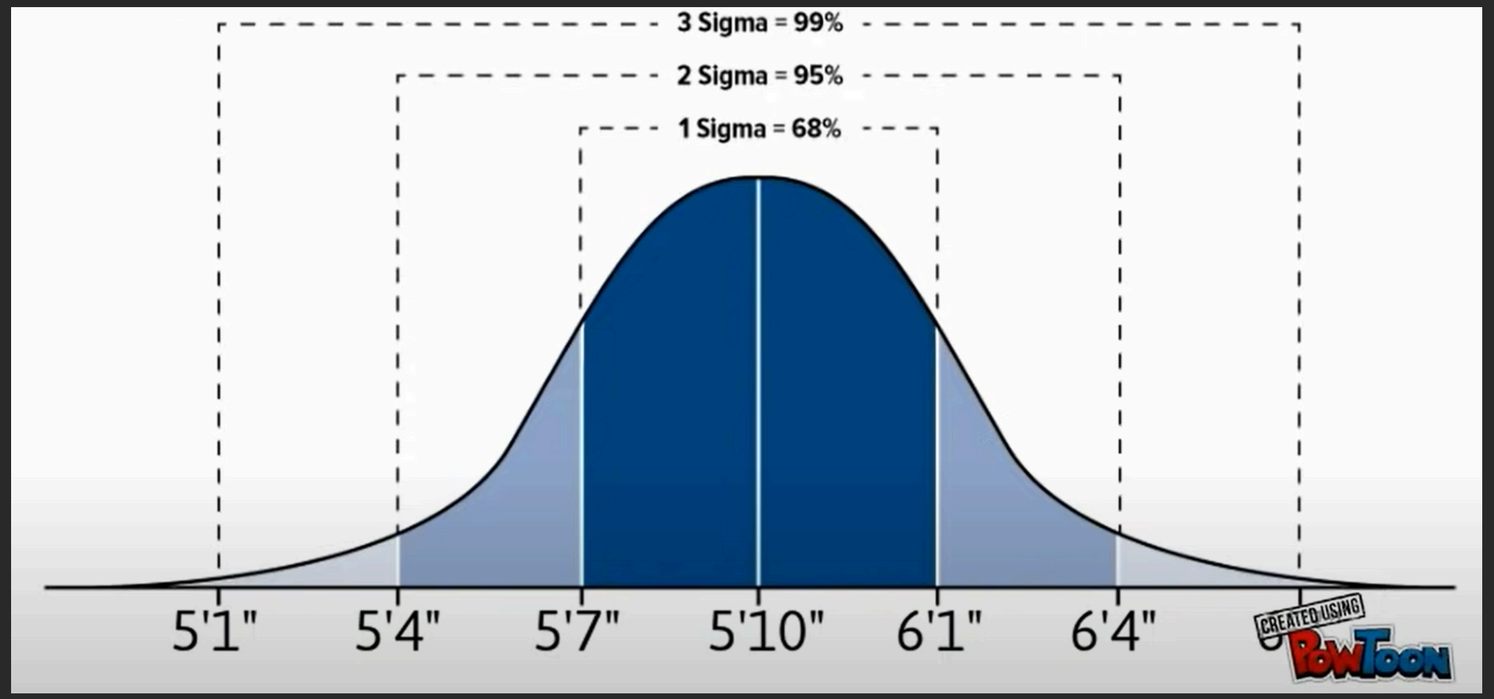
EXAMPLE



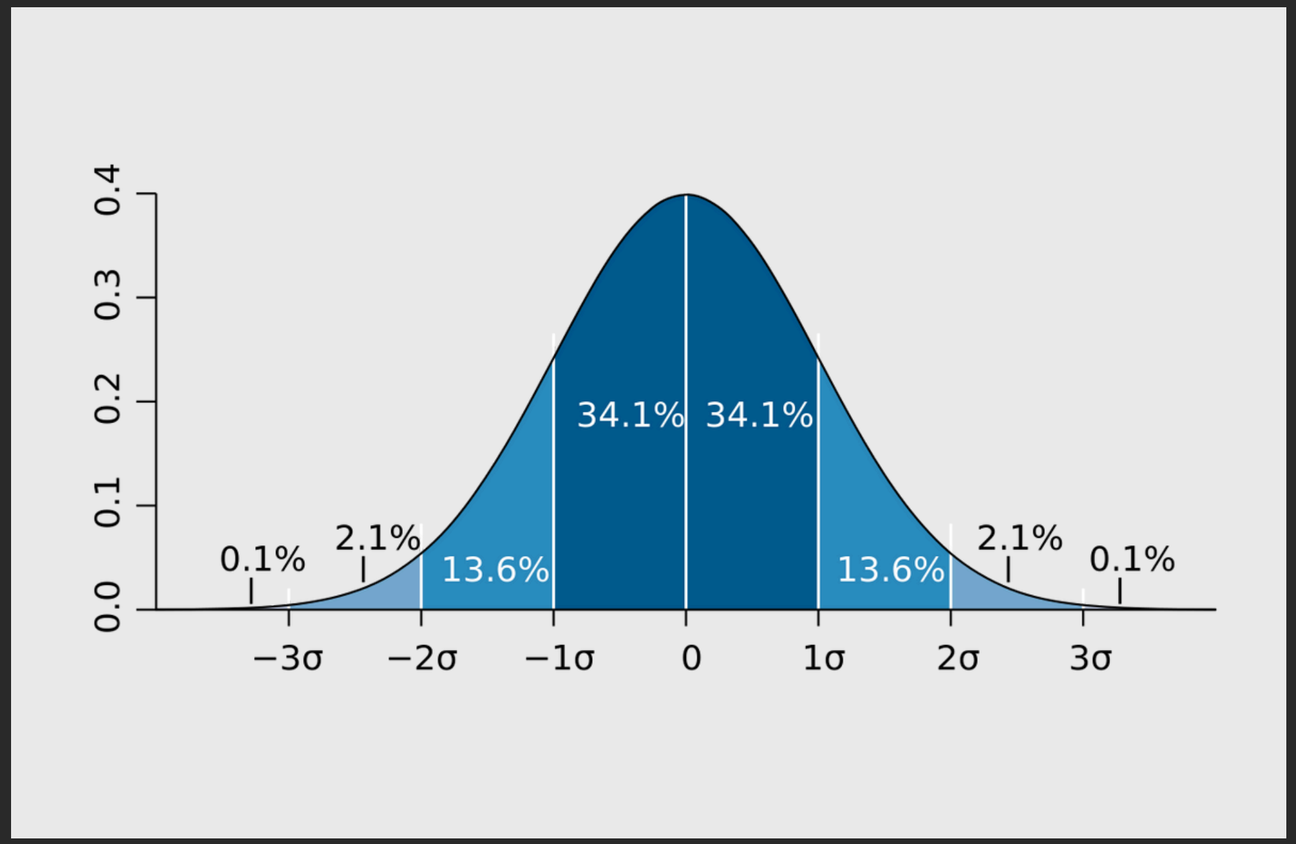
Average height of men in America

EXAMPLE

Only 0.3% of men deviate more than 9 inches of the average height



Average height of men in America



Z-SCORE

The number of standard deviations our data point is away from the mean.

Formula:

$$Z = \frac{x - \mu}{\sigma}$$

EXAMPLE TIME!

Dataset: {4,5,12,23,45,89}

Average: 29.67

STD-dev: 29.9425

Let’s try to find the z-score of 45!

$$\text{Z-Score: } \frac{45 - 29.67}{29.9425} = 0.512$$

So, 45 is 0.512 standard deviations away from the mean. Z-Score = 0.512

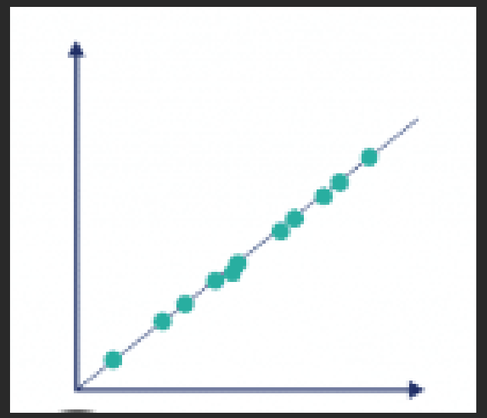
CORRELATION COEFFICIENT

It measures how strong the relationship is between the points.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

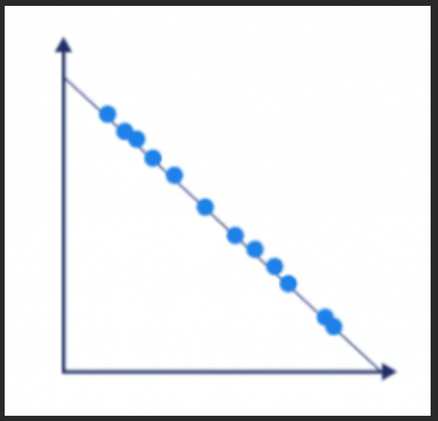
Ranges from -1 to 1

1: Positive Correlation
0: Not a Linear Correlation
-1: Negative Correlation



Perfect Positive

No Correlation



Perfect Negative

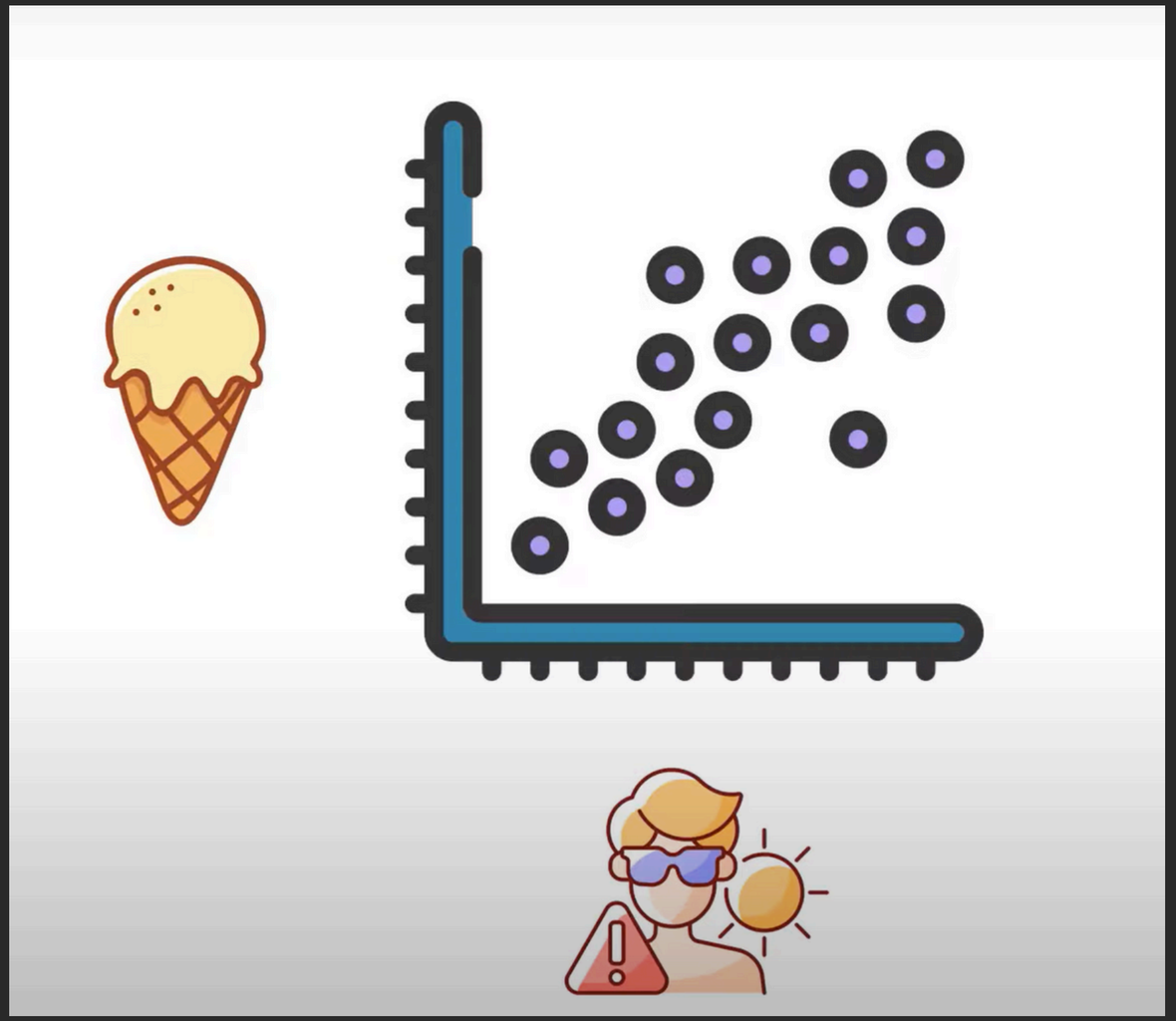
CORRELATION VS. CAUSATION

Correlation is used to understand the relation between variables

However, correlation does not equal causation

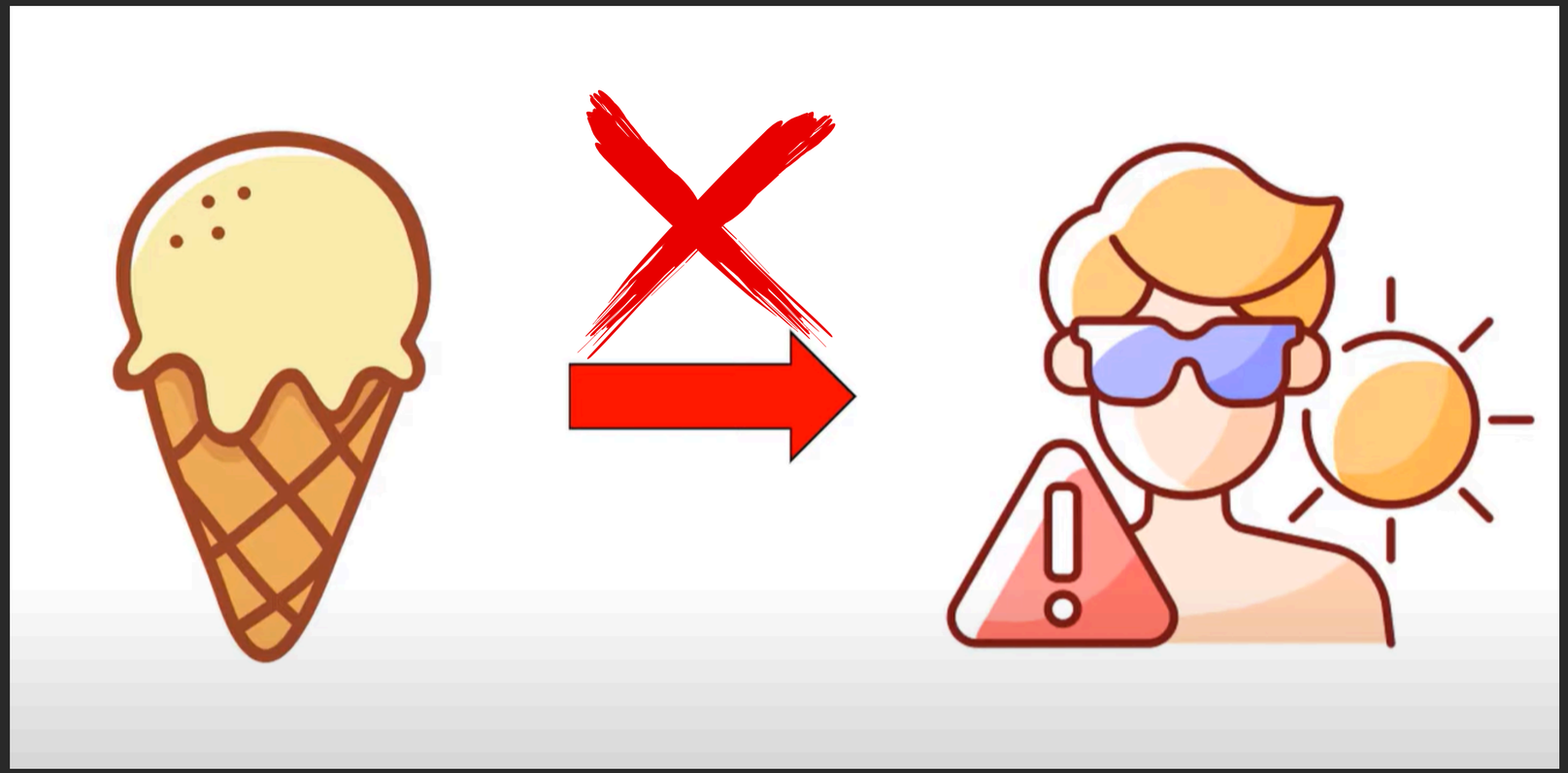
CORRELATION VS. CAUSATION

There exists some correlation between the rate of sunburns and the sale of icecreams

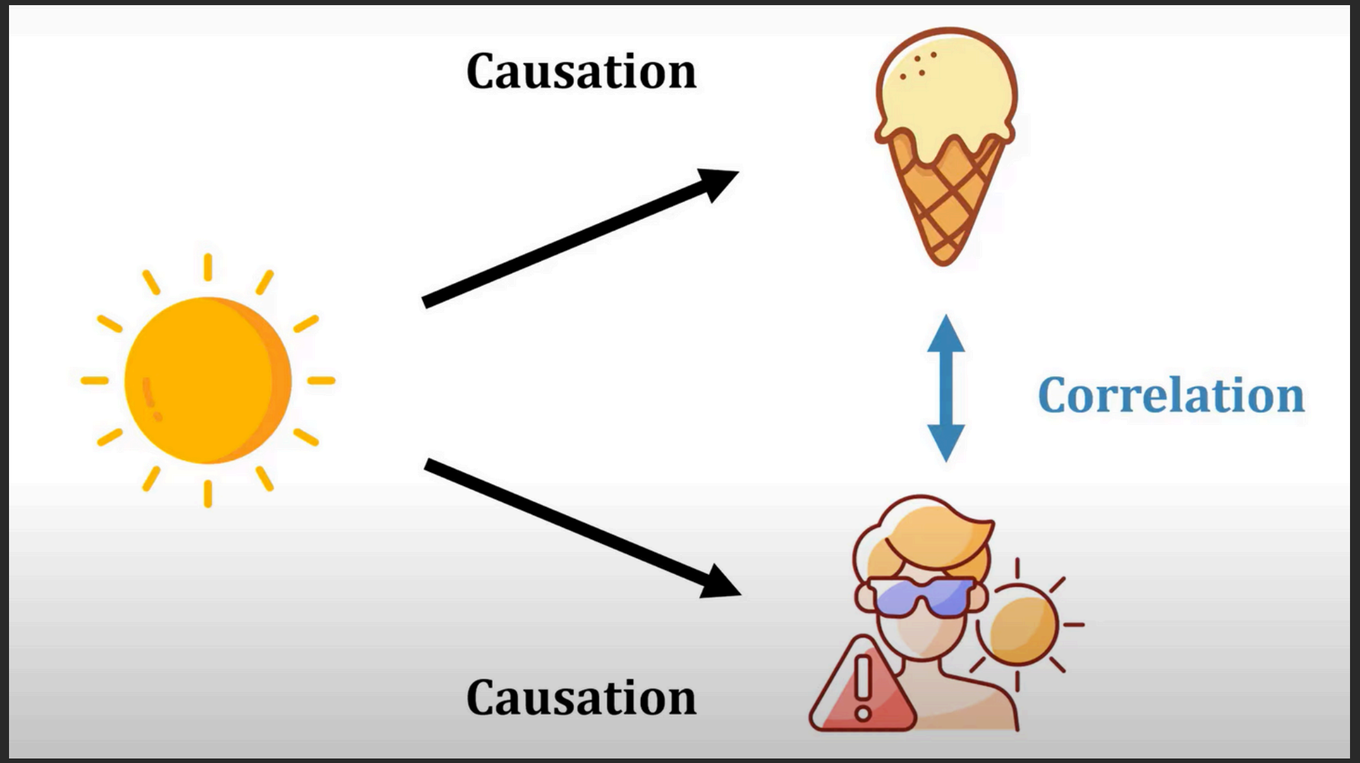


CORRELATION VS. CAUSATION

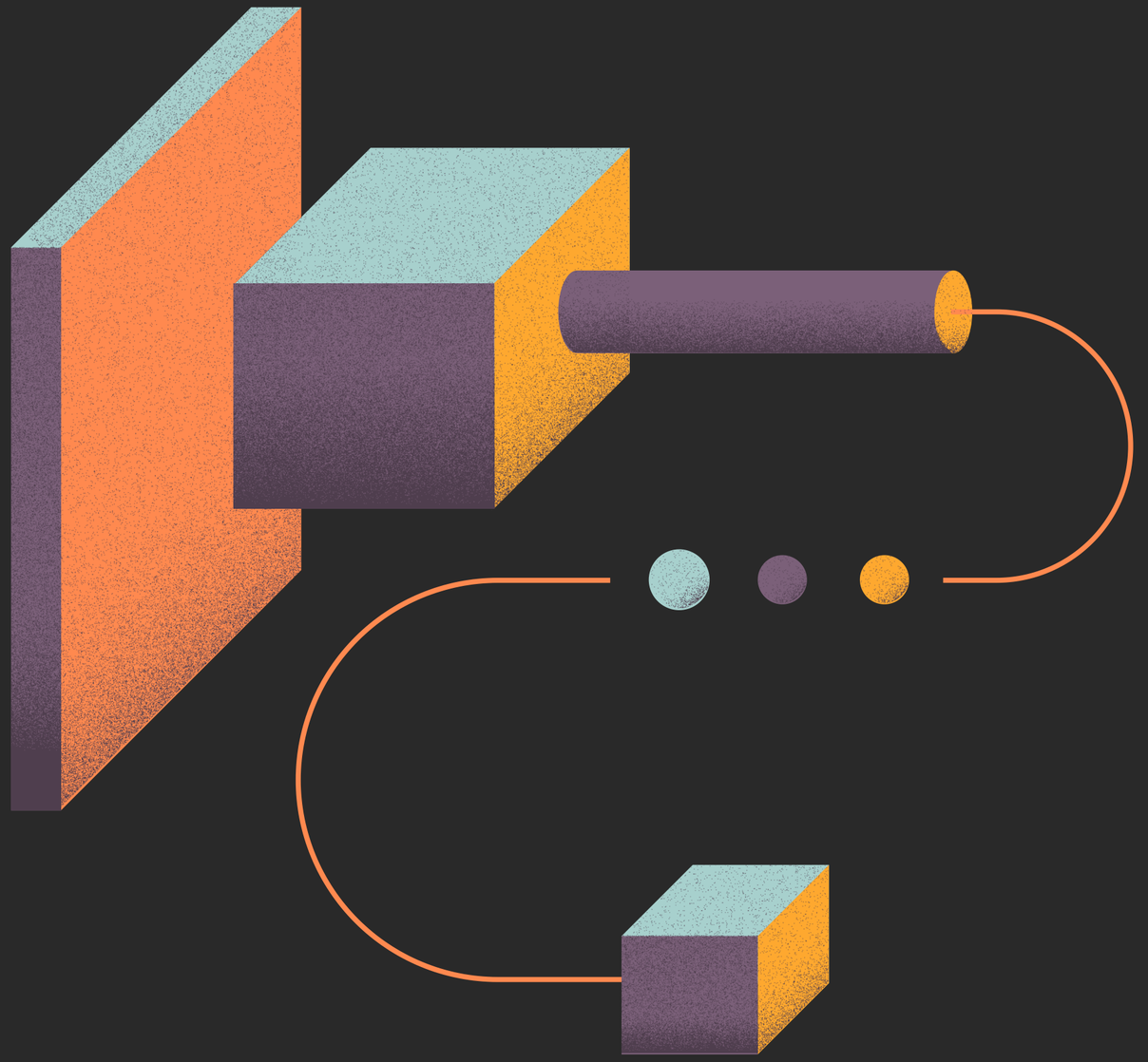
However, icecream does not cause sunburns



CORRELATION VS. CAUSATION



SEE YOU NEXT WEEK!



Data Science Club

COMPANY VISION

Our Vision

Presentations are communication tools that can be used as demonstrations.

Good Management

Presentations are communication tools that can be used as demonstrations.

> **If everyone is moving forward together, then success takes care of itself.**

COMPANY MISSION

Our Mission

Presentations are communication tools that can be used as demonstrations.

Secure Data

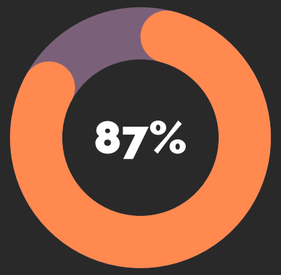
Presentations are communication tools that can be used as demonstrations.

> **No one can whistle a symphony.**
> **It takes a whole orchestra to play it.**

HIGHLIGHT REPORT

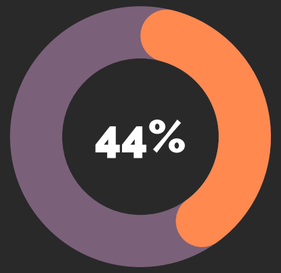
Last Year Report

Presentations are communication tools that can be used as demonstrations.



Big Income

Presentations are communication tools that can be used as demonstrations.



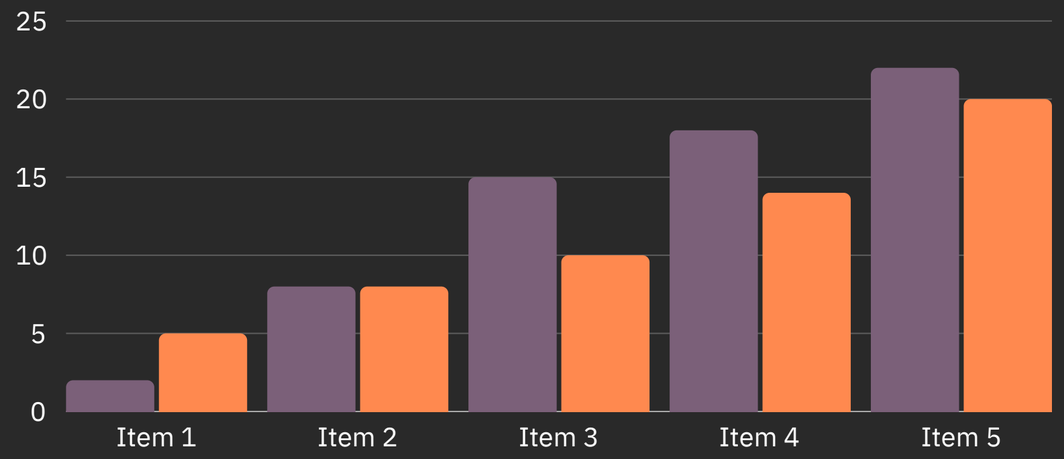
Less Outcome

Presentations are communication tools that can be used as demonstrations.

HIGHLIGHT REPORT

Last Year Report

Presentations are communication tools that can be used as demonstrations.



PROCESS INFOGRAPHIC

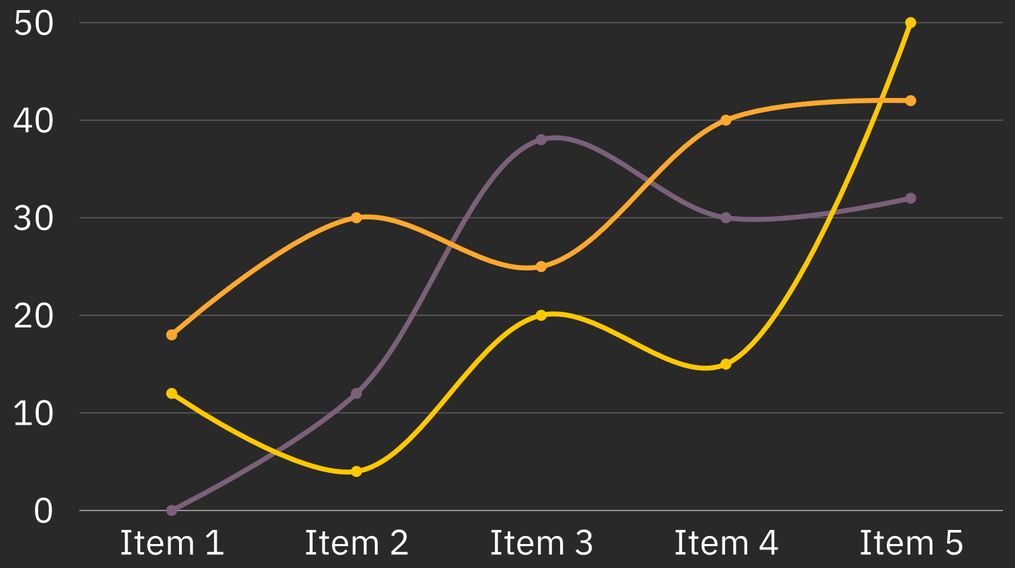
Step One
Communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

Step Two
Communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

Step Four
Communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

Step Thre
Communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

BRAZIL MAP SLIDE



+55%

Market Interest Per Users

Presentations are communication tools that can be used as demonstrations.

SWOT ANALYTICS

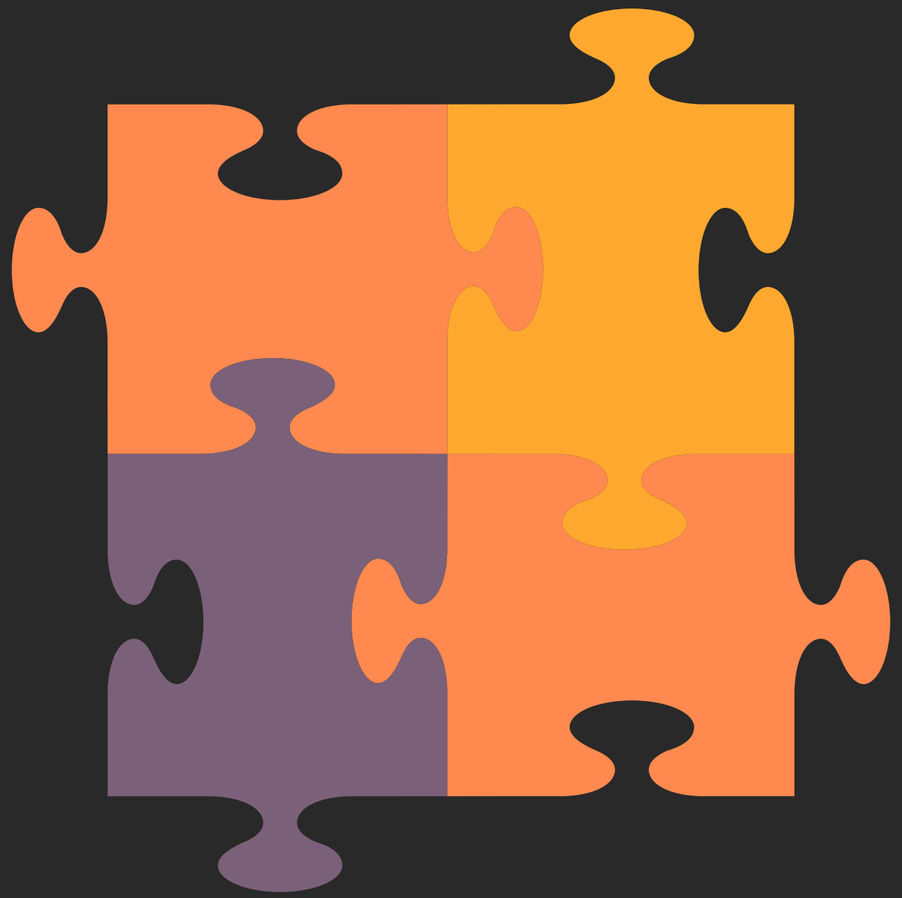
Strength Analytics

Presentations are communication tools that can be used as demonstrations.

- **Aspect One**
- **Aspect Two**
- **Aspect Three**
- **Aspect Four**

Threat Analytics

Communication tools that can be used as demonstrations, lectures, speeches, reports, and more.



Weakness Analytics

Communication tools that can be used as demonstrations, lectures, speeches, reports, and more.

Opportunity Analytics

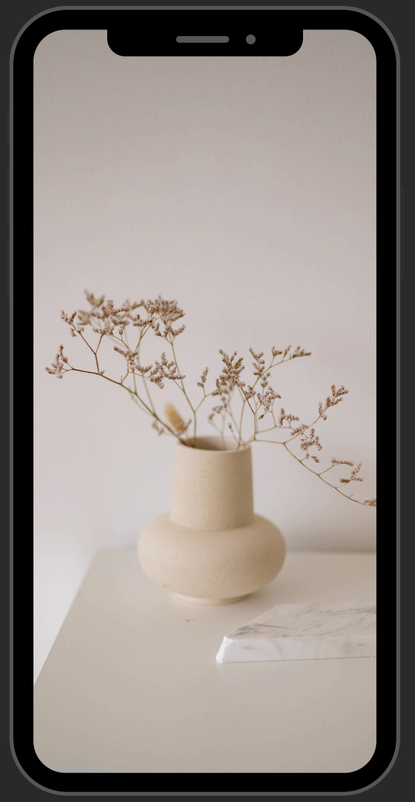
Presentations are communication tools that can be used as demonstrations.

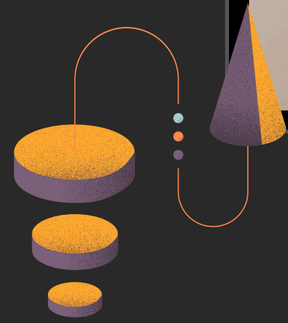
- **Aspect Three**
- **Aspect Four**
- **Aspect One**
- **Aspect Two**

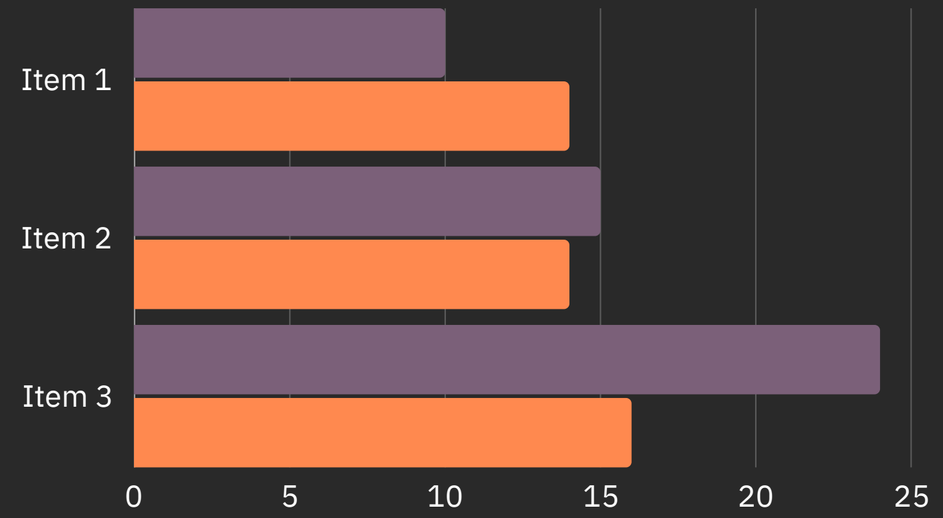
MOCKUP SLIDE

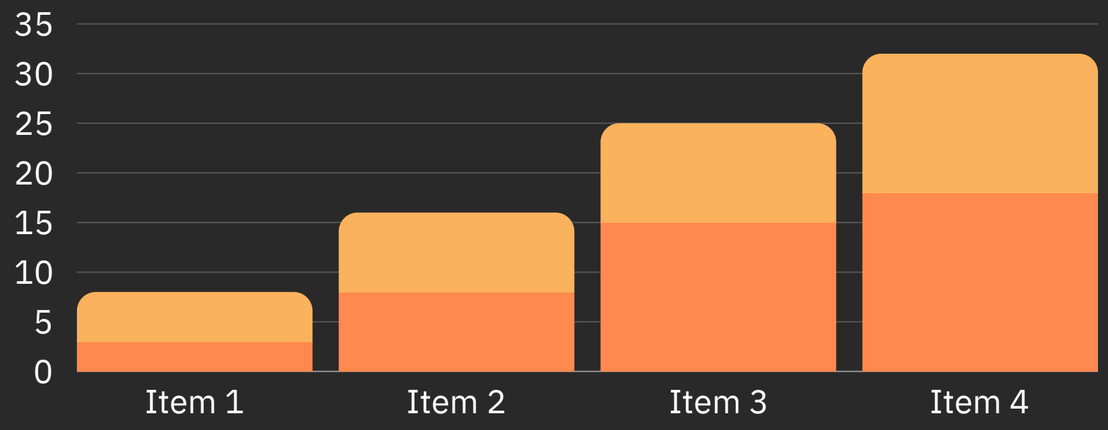
Feature Explaining

Presentations are communication tools that can be used as demonstrations.









MOCKUP SLIDE

Feature One

Presentations are communication tools that can be used as demonstrations

Feature Two

Presentations are communication tools that can be used as demonstrations