# Summer of Code
# **Artificial Intelligence**
## (Machine Learning & Deep Learning)

Instructor
**Wajahat Ullah**
**-** *Research Assistant* (DIP Lab)

Duration
**03 Months**
(September – November)

# Day 04 – Descriptive Statistics (Data and Its Types)

**Objectives:**

- ❖ What is Descriptive Statistics?
- ❖ What is Data?
- ❖ Different Types of Data
- ❖ Measure of Central Tendency

# Statistics

*[stə-'ti-stiks]*

A branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data.
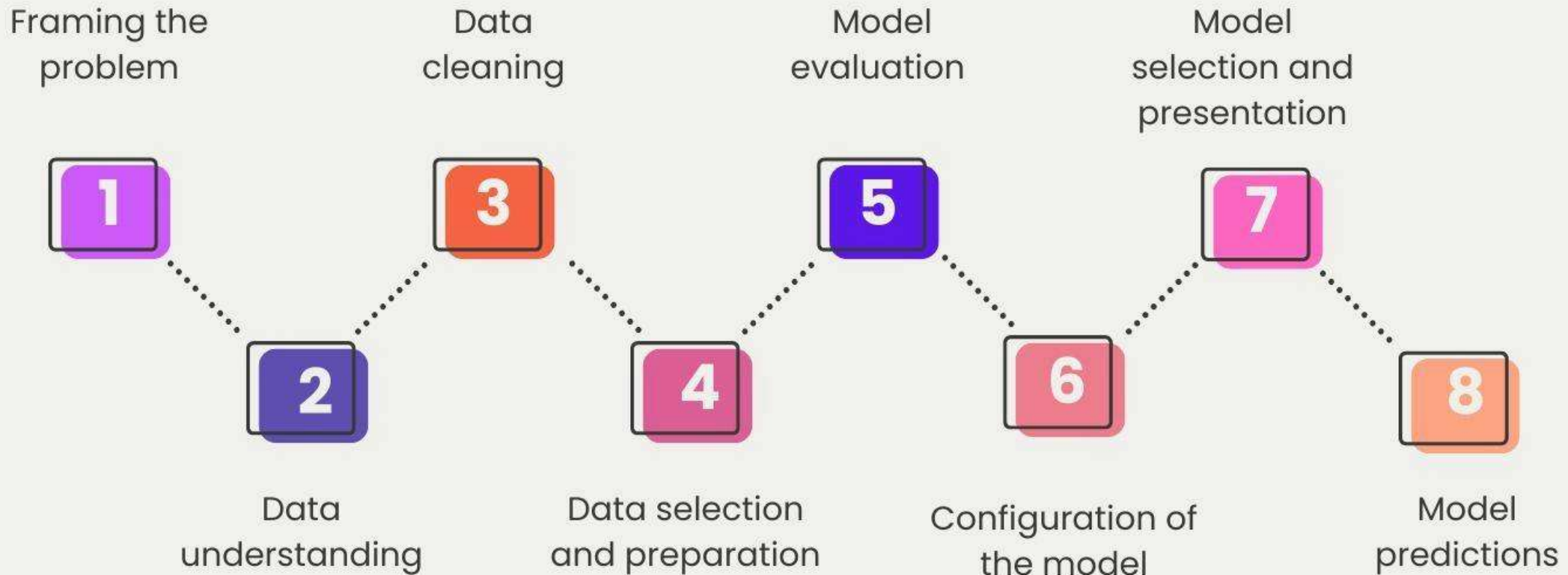
# Descriptive Statistics

*[di-ˈskrip-tiv stə-ˈti-stiks]*

Statistics that summarize or describe features of a data set, such as its central tendency or dispersion.

# How Statistics is Used in Machine Learning

Framing the problem

**1**

Data understanding

**2**

Data cleaning

**3**

Data selection and preparation

**4**

Model evaluation

**5**

Configuration of the model

**6**

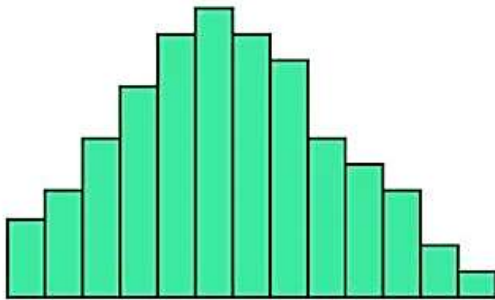Model selection and presentation

**7**

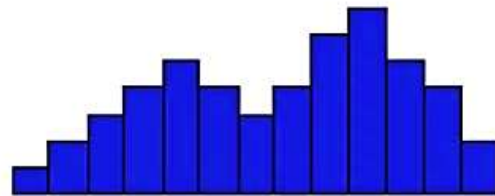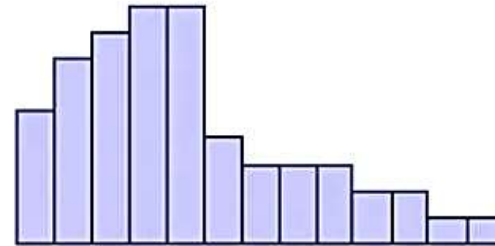Model predictions

**8**

# Types of Data
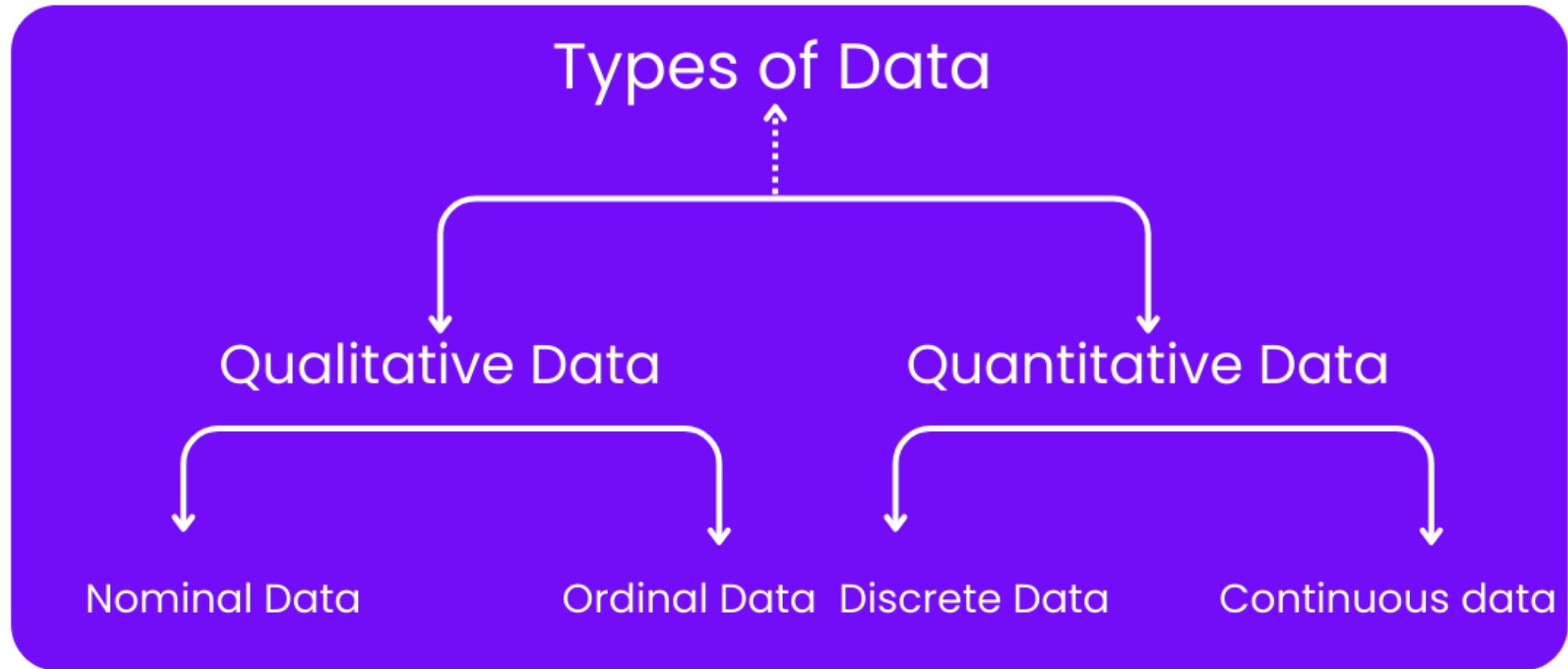
# Data Distribution

Data distribution describes how values in a dataset are spread or clustered across a range. It reveals the nature of the data, providing insight into its central tendency, variability, and shape.



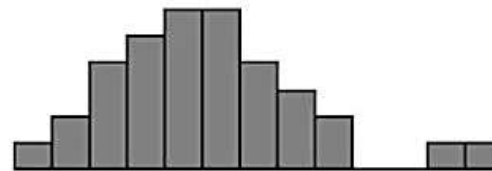The shape has a bell shape.
It is symmetric.

The shape has two humps.
It is bimodal.

The shape has a long tail.
It is not symmetric.

The shape is flat.

There are one or more outliers.

# Qualitative Data

- Qualitative data refers to non-numerical information that describes characteristics.

- Focuses on words, meanings, opinions, feelings, and descriptions.

- Gathered through interviews, open-ended surveys, focus groups, observations etc.

- Analyzed through categorization, or coding to identify patterns and insights



Customer satisfaction

Product categories

New vs. returning visitors

User demographics
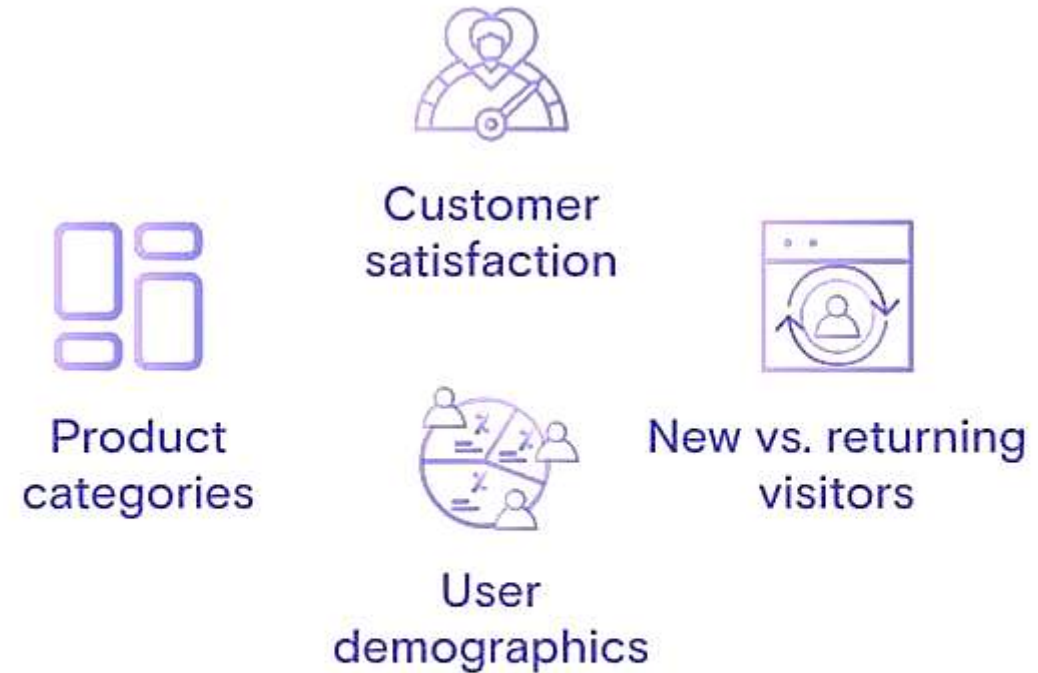
Qualitative data consists of categorical, descriptive data
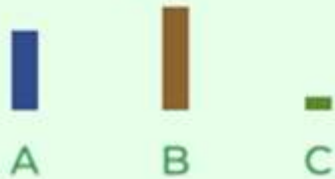
# Qualitative Data

# Qualitative Data



## Ordinal Data

Ordinal data classifies variables into categories which have a natural order or rank.

### Examples

**School grades**

A    B    C

**Educational level**

Bachelor's   Master's   PhD

**Seniority level**

Junior   Mid   Senior

**How is Ordinal Data Analyzed ?**

**Descriptive statistics:** Frequency Distribution, mode, median, and range

**Non-parametric** statistical tests

# Quantitative Data

- Quantitative data refers to numerical information that can be measured, counted, and expressed in numbers.

- Focuses on quantities, amounts, and measurable values.

- Gathered through surveys with close-ended questions, experiments, observations.

- analyzed using statistical methods, graphs, and mathematical models to identify patterns and trends
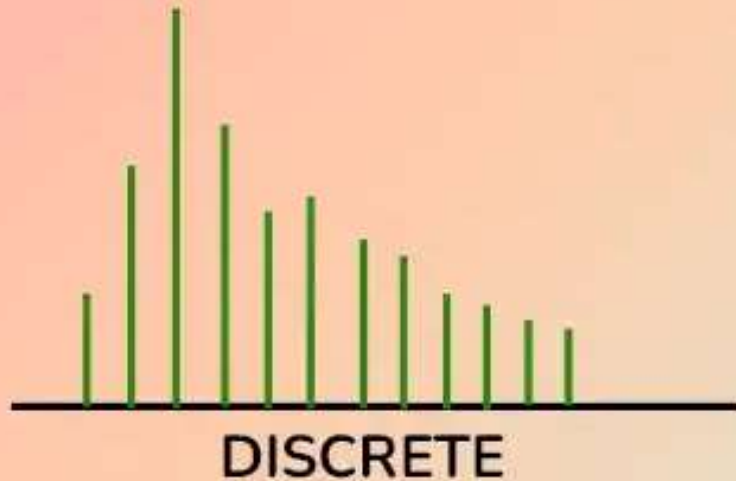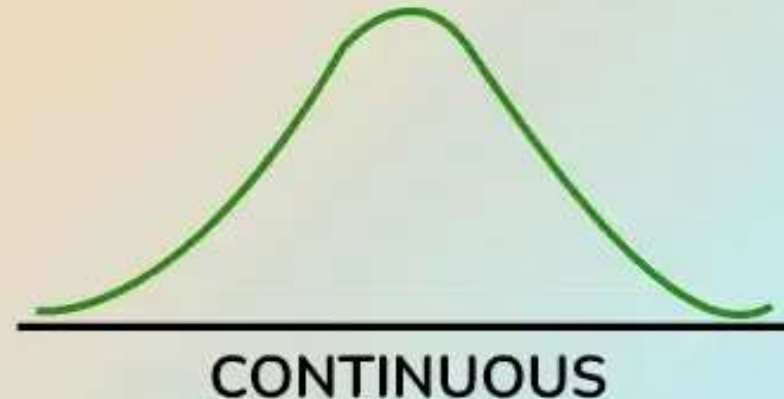


Sales numbers
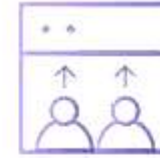
Visitor numbers

Ad clicks

Product failure rates

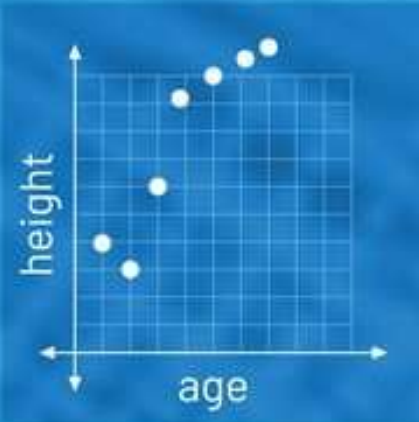Quantitative data consists of numerical, measurable values

# Quantitative Data



**DISCRETE**
- Obtained by counting values such as integers 0,1,2,3...
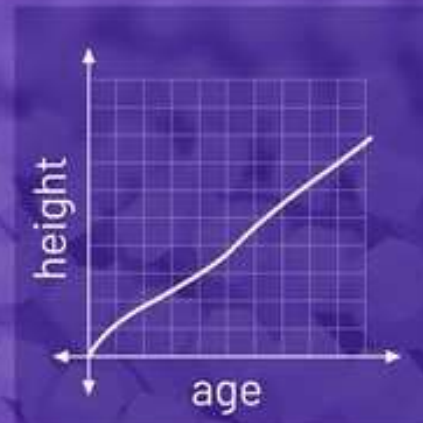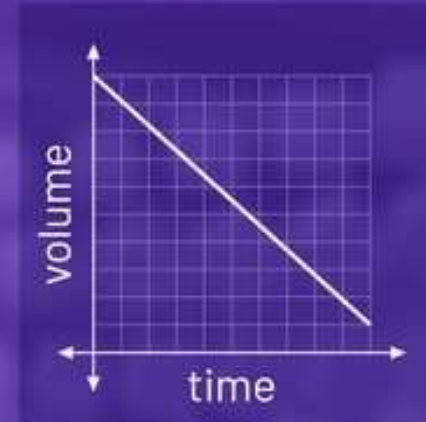- Example: Your score in this upcoming mid-term exams

**CONTINUOUS**
- Obtained from the data that can take infinitely many values
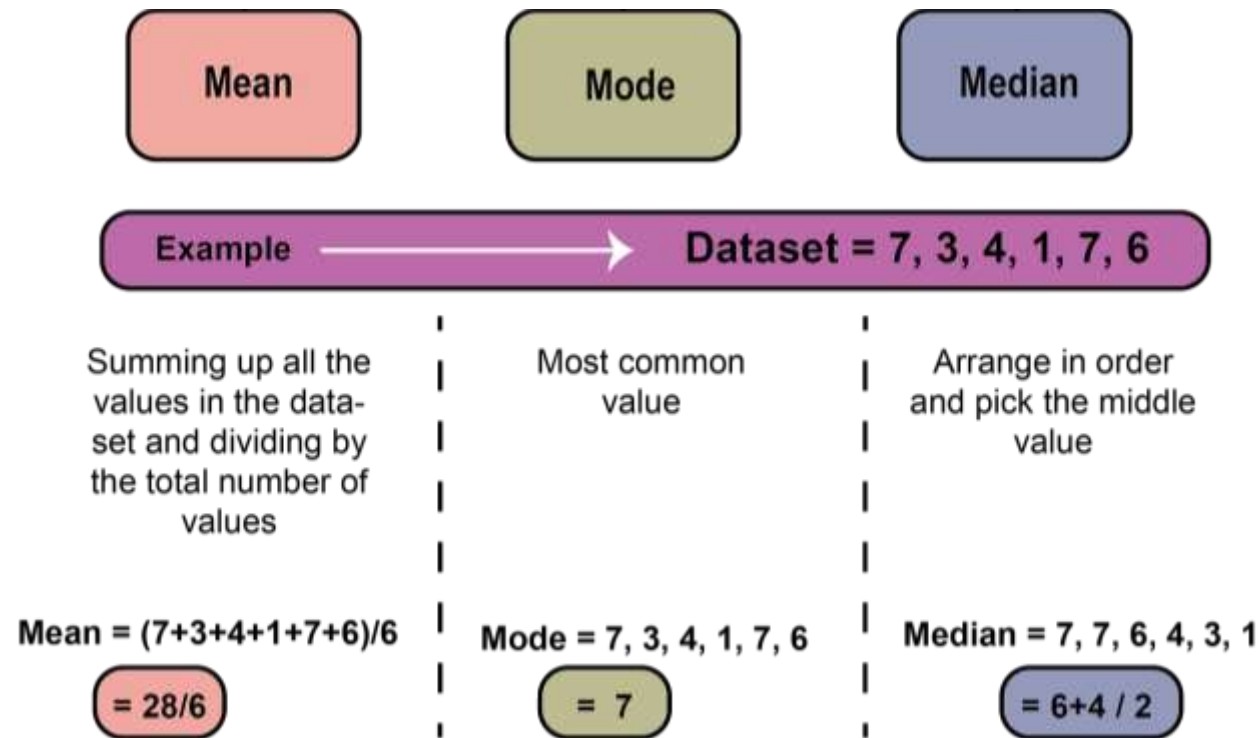- Example: The expected lifetime of a new light bulb

# Quantitative Data

# Measures of Central Tendency

These are single summary values that describe the central location of a dataset. The mean is the average value, the median is the middle number in an ordered dataset, and the mode is the most frequent value.
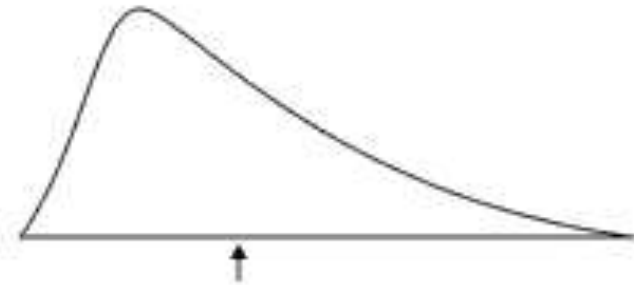
| Mean | Mode | Median |
|------|------|--------|

Example ———————————————➤ Dataset = 7, 3, 4, 1, 7, 6

| Summing up all the values in the data-set and dividing by the total number of values | Most common value | Arrange in order and pick the middle value |
|---|---|---|
| Mean = (7+3+4+1+7+6)/6 | Mode = 7, 3, 4, 1, 7, 6 | Median = 7, 7, 6, 4, 3, 1 |
| = 28/6 | = 7 | = 6+4 / 2 |

# Mean or Average

- The mathematical average of a set of two or more numbers is called its mean.

- Mathematically,

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$
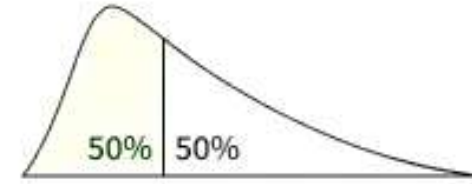
Mean is the average or norm.

# Median

- The median is the middle value in a dataset that has been arranged in ascending or descending order.

- When the number of elements is odd:

$$median = \left(\frac{n+1}{2}\right)^{th}$$

- When the number of elements is even:

$$median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

50% | 50%

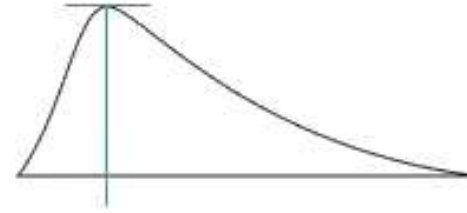Meadian is the middle value.

5, 13, 9, 7, 1, 9, 2, 9, and 11

put in ascending order

1, 2, 5, 7, 9, 9, 9, 11, 13
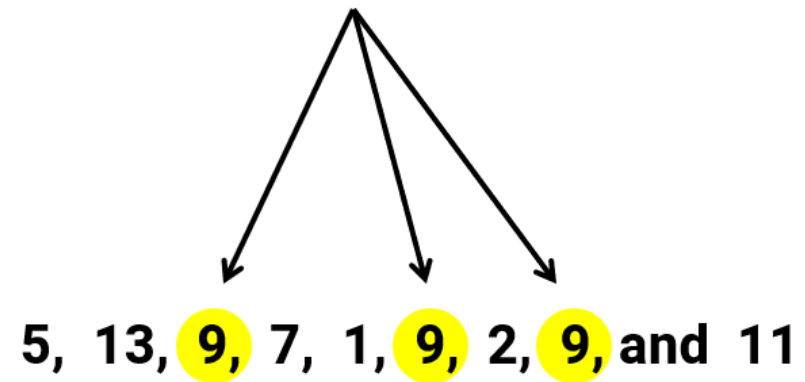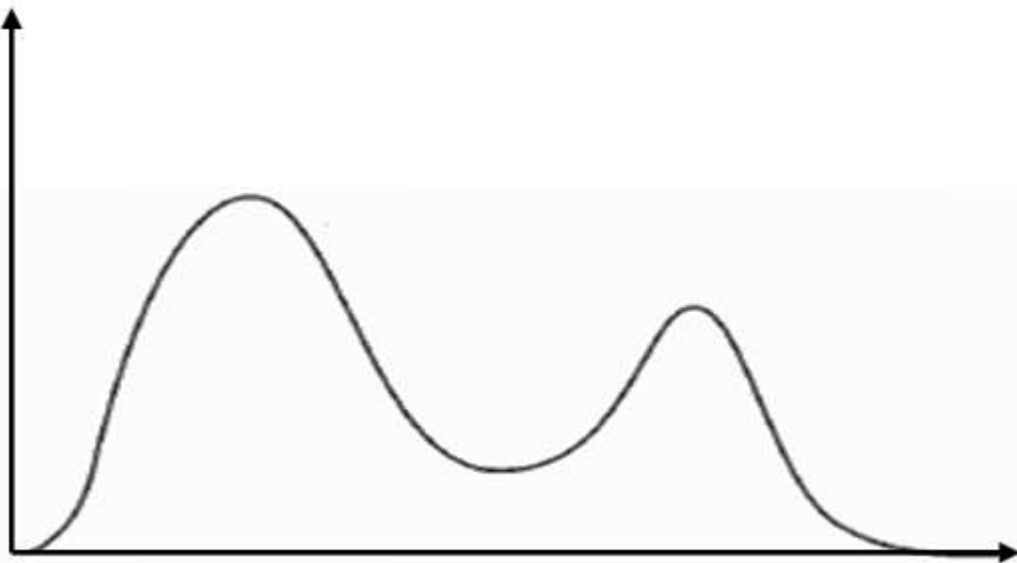
Median (middle value)

# Mode

- The mode is the most frequently occurring value in a dataset.

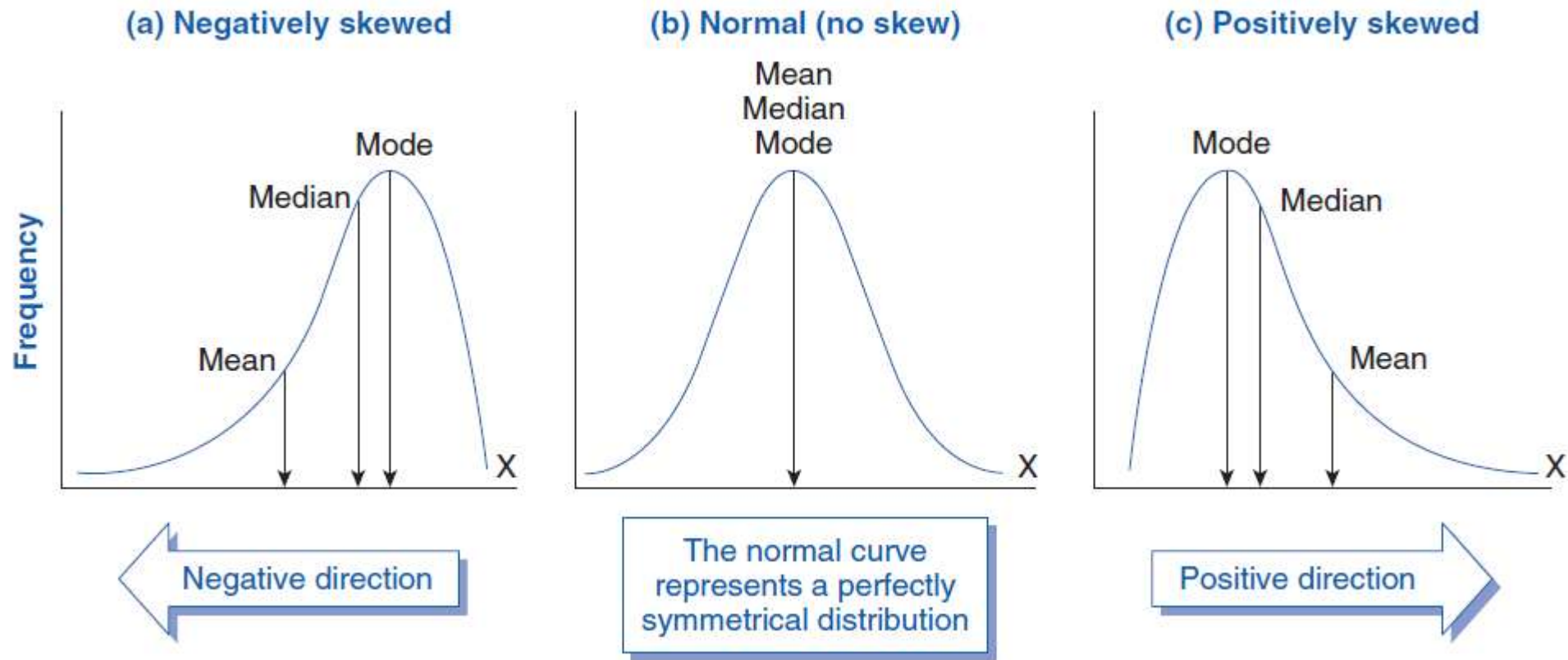- There can be more than one mode, or no mode at all.

Mode is the most frequent value.

Shows up the most!

5, 13, **9,** 7, 1, **9,** 2, **9,** and 11

Mode = **9**

# Relationship Between Mean, Median, and Model

# Happy Coding