Data Visualization Techniques

Week 2
Lecture 1
Spring 2025
Rida Maryam

Why is Data Visualization Important?

- Helps in identifying patterns and trends.
- Makes data easier to understand.
- Supports decision-making in business, research, and daily life.

Example:

A business owner analyzes sales using a graphs instead of raw numbers.

Data Visualization Techniques

- Tables
- Graphs
  - Bar Chart
  - Pie Chart
  - Histogram
  - Line Chart
  - Scatter Plot

Data Visualization

1. **Table:** The very simplest way to present or visualize a dataset is to produce a table

For example:

Index	Net worth	Index	Taste score	Index	Taste score
1	100, 360	1	12.3	11	34.9
2	109, 770	2	20.9	12	57.2
3	96, 860	3	39	13	0.7
4	97, 860	4	47.9	14	25.9
5	108, 930	5	5.6	15	54.9
6	124, 330	6	25.9	16	40.9
7	101, 300	7	37.3	17	15.9
8	112, 710	8	21.9	18	6.4
9	106, 740	9	18.1	19	18
10	120, 170	10	21	20	38.9

Drawback of table

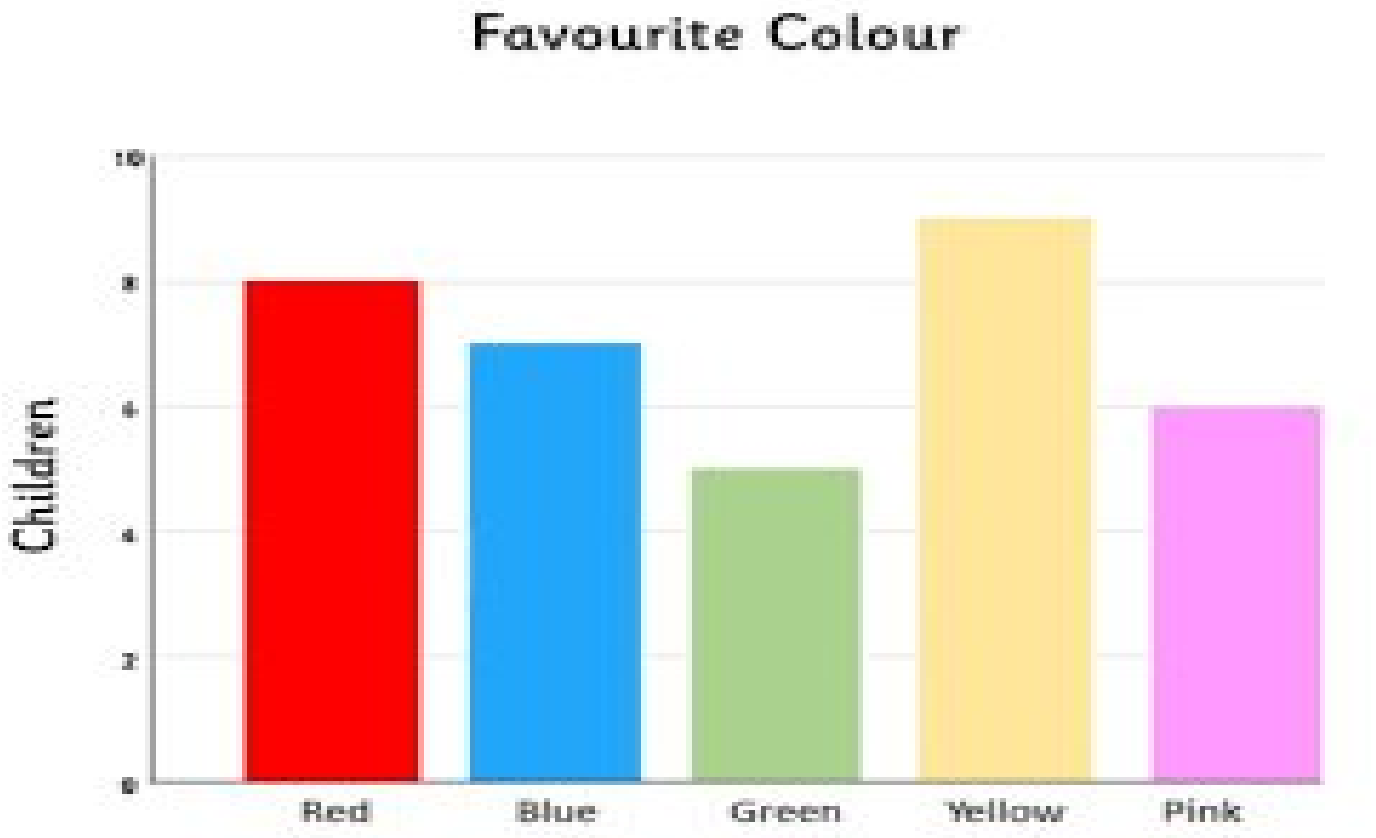
Tables can be helpful, but aren’t much use for large datasets, because it is difficult to get any sense of what the data means from a table.

What is a Bar Chart?

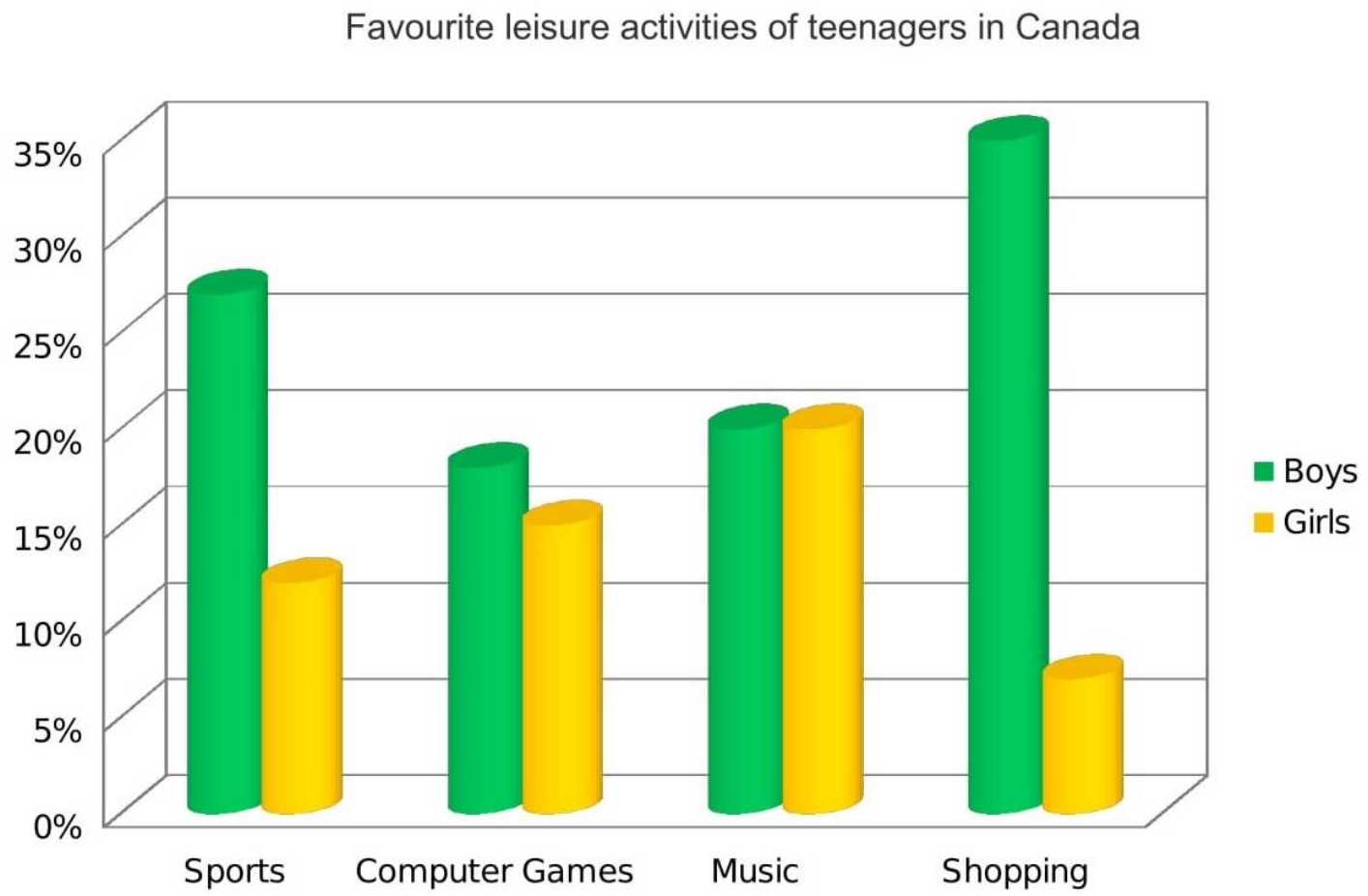
- A graphical representation of categorical data using rectangular bars.
- The height of the bar represents the frequency of each category.
- Bars can be vertical or horizontal.

Example:
Favorite color survey of children.
Favorite leisure activity of teenagers
Favorite sports of students
Number of population over the years

Favorite color survey of children.



Favorite leisure activity of teenagers



Key Features of a Bar Chart

- ✔ Bars can be arranged in any order.
- ✔ Bars have equal spacing between them.
- ✔ Used for categorical (non-numeric) data.
- ✔ Colors and labels help in interpretation.

What is a Pie Chart?

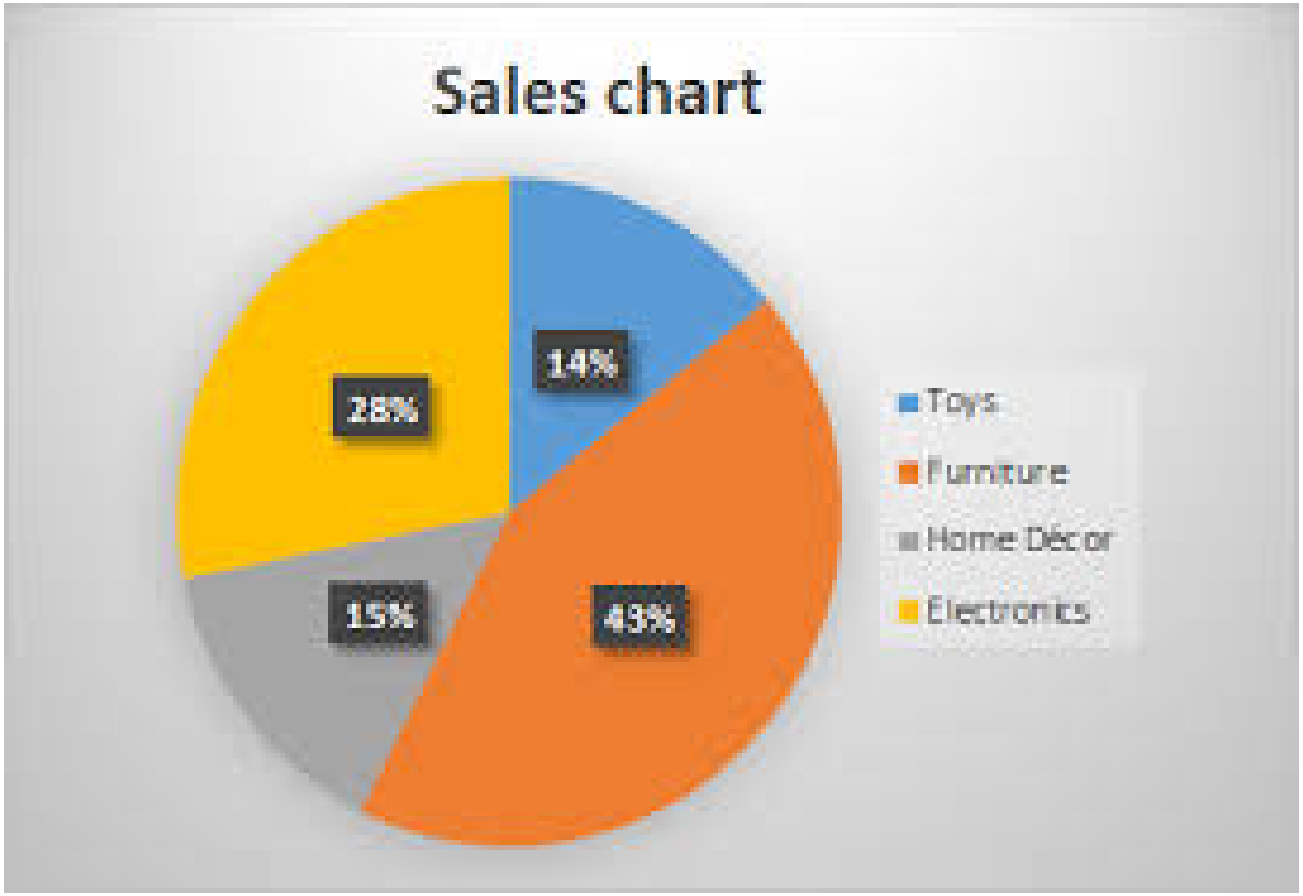
- Type of graph in which a circle is divided into sectors that each represent a proportion of the whole.
- Pie slices of the chart show the relative size of the data.
- Best for showing percentages or proportions

Example:

- Percentage of recommended Diet
- Sales chart
- market share of companies
- Favorite beverages
- Percentage of students enrolled

Percentage of recommended Diet

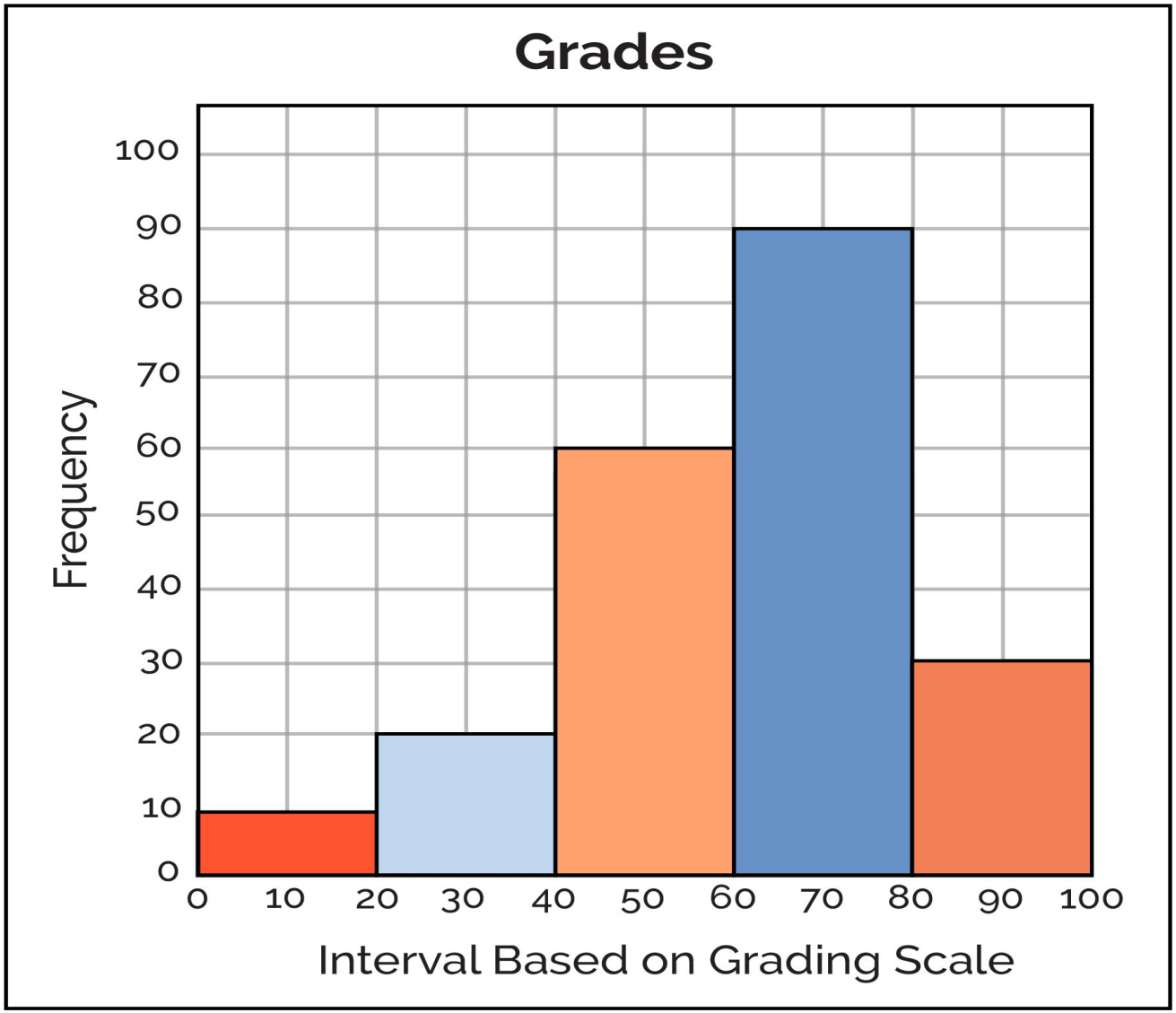
Sales chart


Sales chart
14%
43%
15%
28%
Toys
Furniture
Home Décor
Electronics

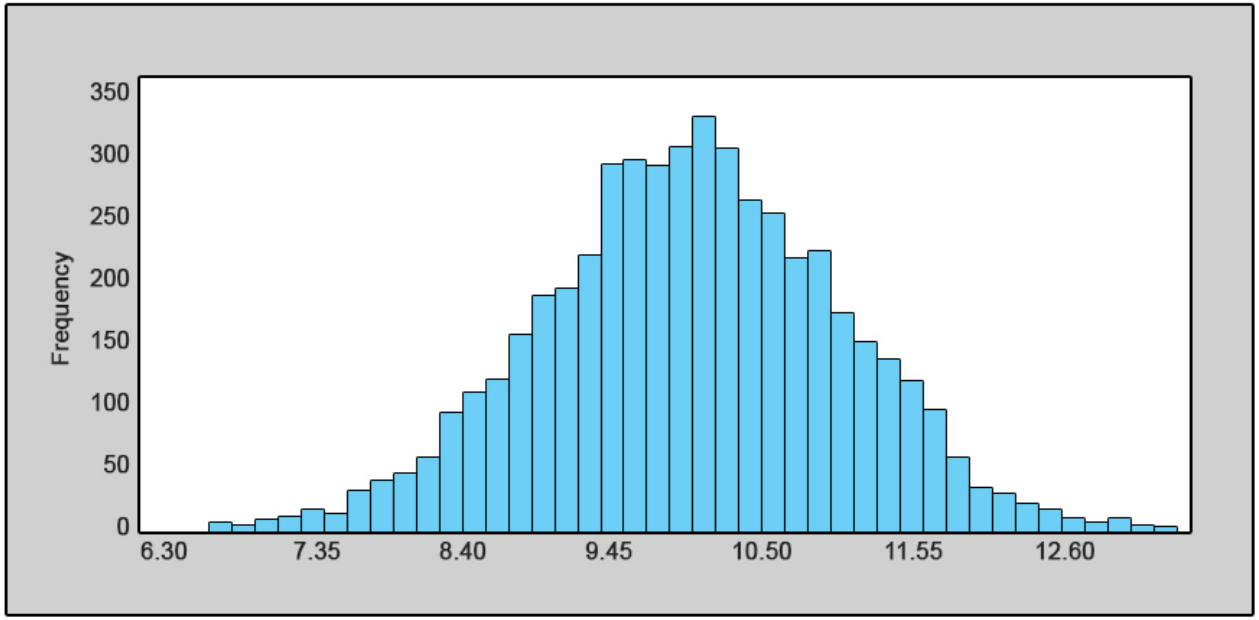
What is a Histogram?

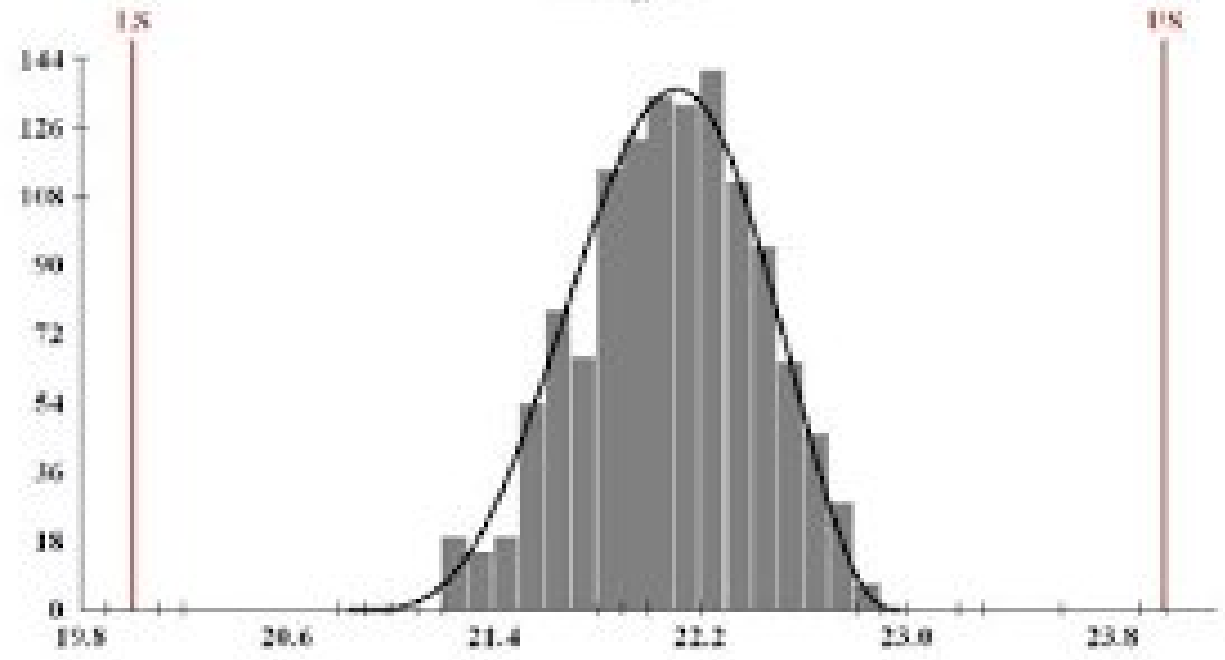
- A graphical representation of numerical (continuous) data.
- It shows how frequently data falls into different intervals (bins).
- Unlike a bar chart, histograms have no gaps between bars.

Example:
A teacher records exam scores into intervals.
(A histogram can show how students are distributed across different score ranges.)

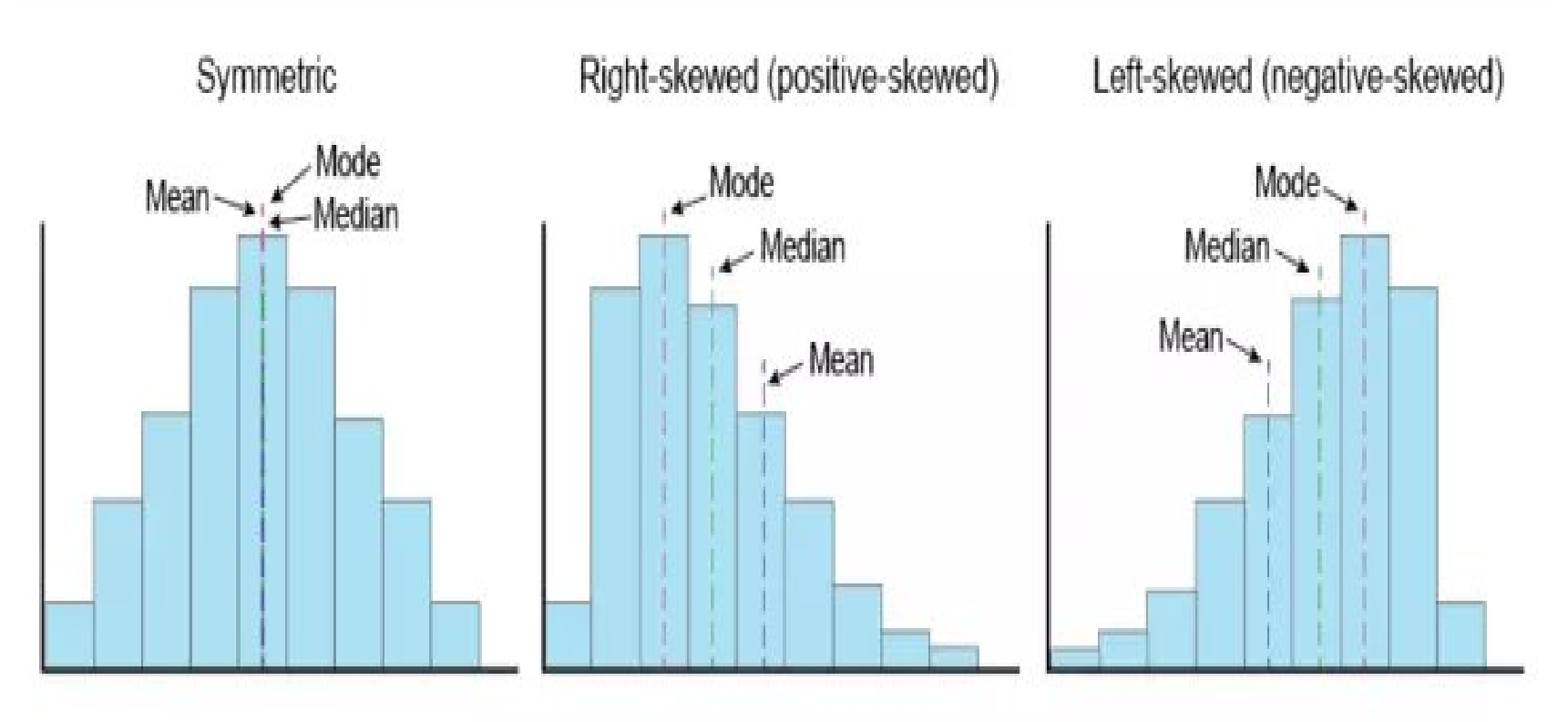

Grades
Frequency
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
Interval Based on Grading Scale


Age Group Distribution
No. of Possible Consumer
50
45
40
35
30
25
20
15
10
5
0
05-09
10-14
15-19
20-24
25-29
Age Groups


Frequency
350
300
250
200
150
100
50
0
6.30
7.35
8.40
9.45
10.50
11.55
12.60


LSL
USL
144
126
108
90
72
54
36
18
0
19.8
20.6
21.4
22.2
23.0
23.8

Skewness


Symmetric
Mean
Mode
Median
Right-skewed (positive-skewed)
Mode
Median
Mean
Left-skewed (negative-skewed)
Mode
Median
Mean

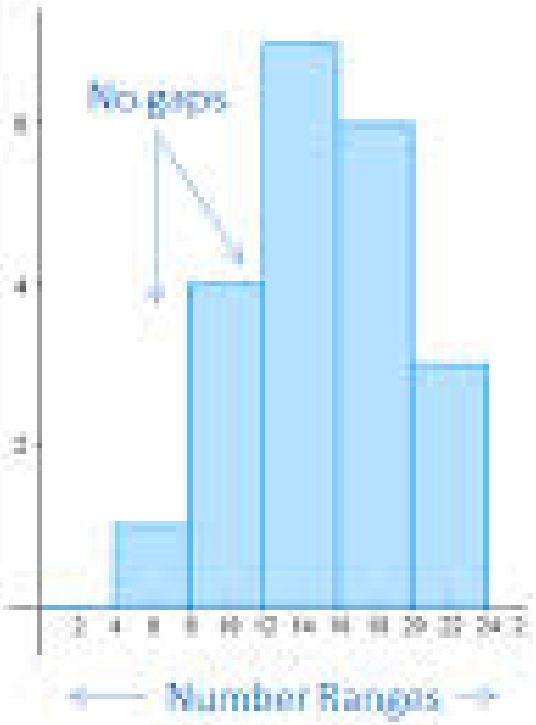
Key Features of a Histogram

✔ No gaps between bars (continuous data).
✔ Used for numerical data divided into intervals.
✔ The area of bars represents frequency, not just height.
✔ Helps in understanding data distribution.

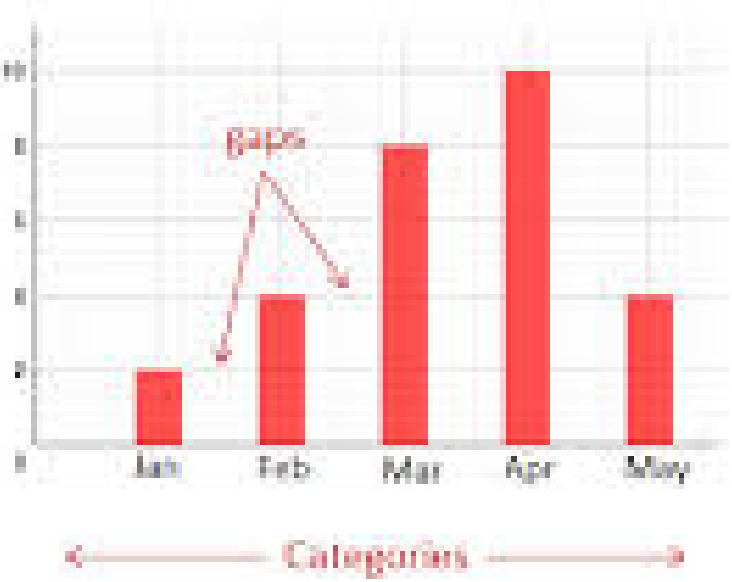
Difference Between Bar Charts and Histograms

Feature	Bar Chart	Histogram
Data Type	Categorical	Numerical (continuous)
Gaps Between Bars	Yes	No
Order of Bars	Can be changed	Fixed (based on bins)
Purpose	Compare categories	Show data distribution

Histogram vs. Bar Chart



Histogram



Bar Chart

Frequency and Relative Frequency

- **Frequency**: Count of occurrences in each category/bin.
- **Relative Frequency**: Percentage of total occurrences.

Relative Frequency=Frequency/Total Count×100

- **Example:**
  If 20 students like football out of 100 students:
  Relative Frequency=20/100×100=20%

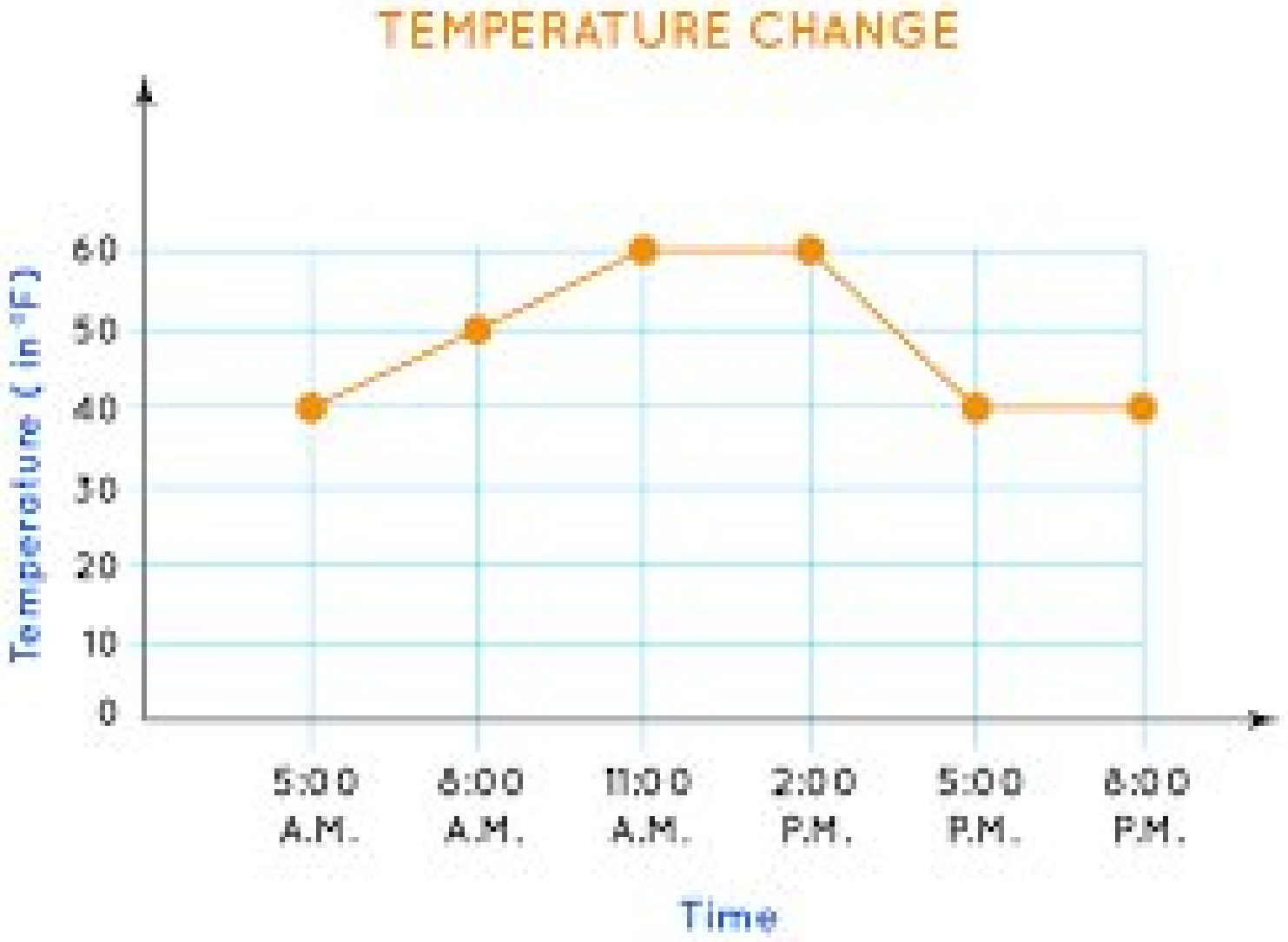
- Frequency and relative frequency help in comparing data points.

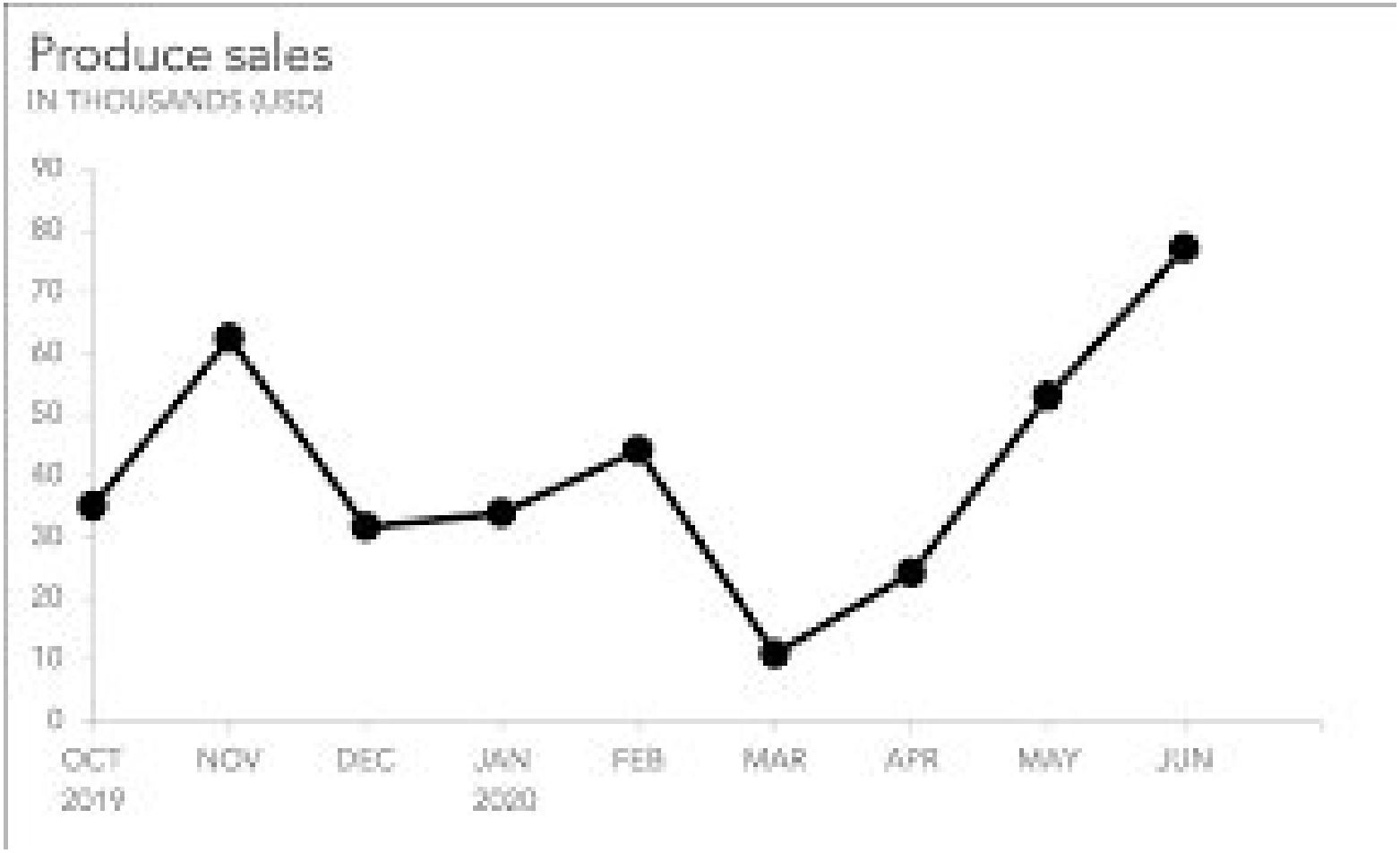
What is a Line Chart?

- Best for visualizing trends and changes over time, connecting data points with lines.
- Also known as line graph or line plot.

Examples:

- Stock prices
- Temperature changes
- Sales trends
- Consumer interest
- Average rainfall over the years


TEMPERATURE CHANGE
Temperature (in °F)
60
50
40
30
20
10
0
5:00 A.M.
8:00 A.M.
11:00 A.M.
2:00 P.M.
5:00 P.M.
8:00 P.M.
Time

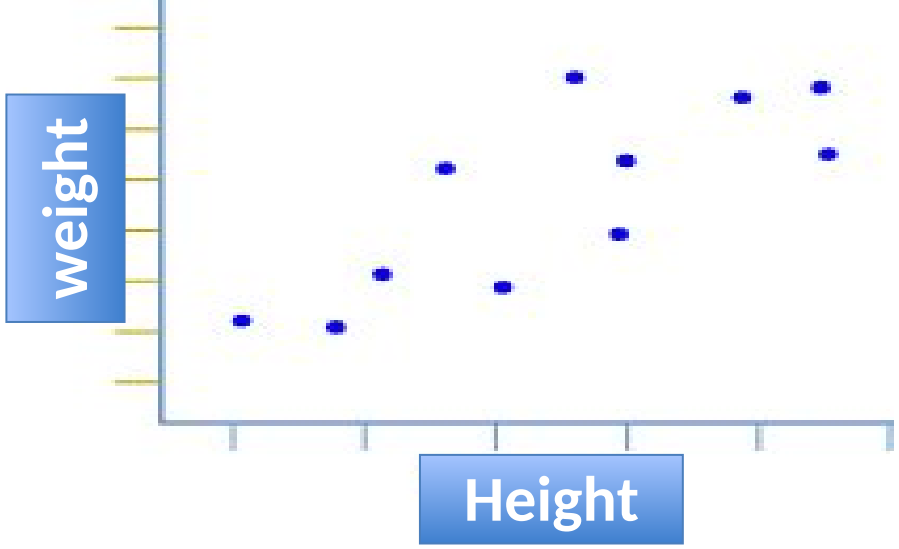

Produce sales
IN THOUSANDS (USD)
90
80
70
60
50
40
30
20
10
0
OCT
2019
NOV
DEC
JAN
2020
FEB
MAR
APR
MAY
JUN

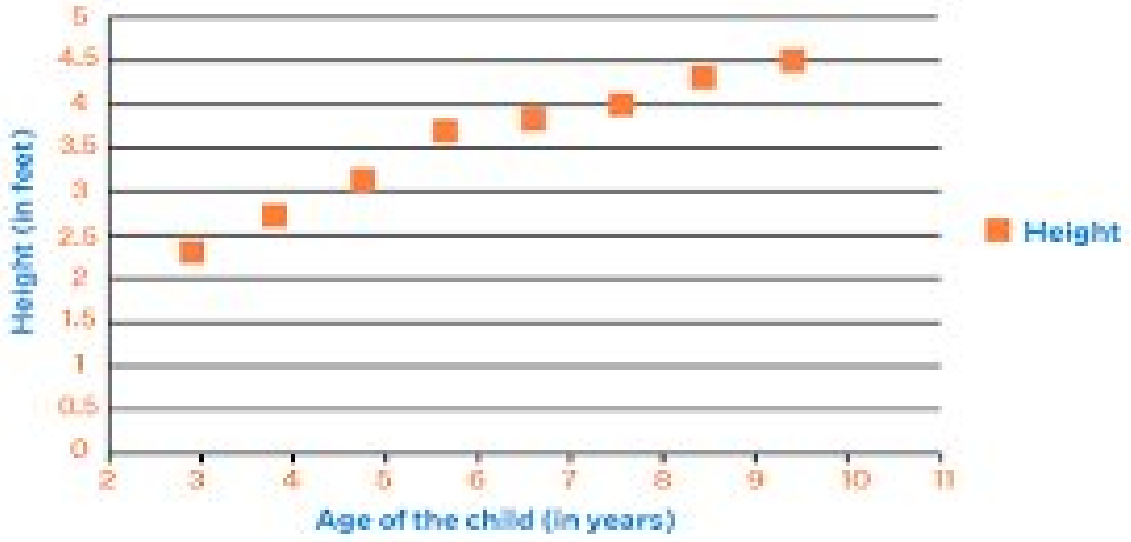
What is a Scatter Plot?

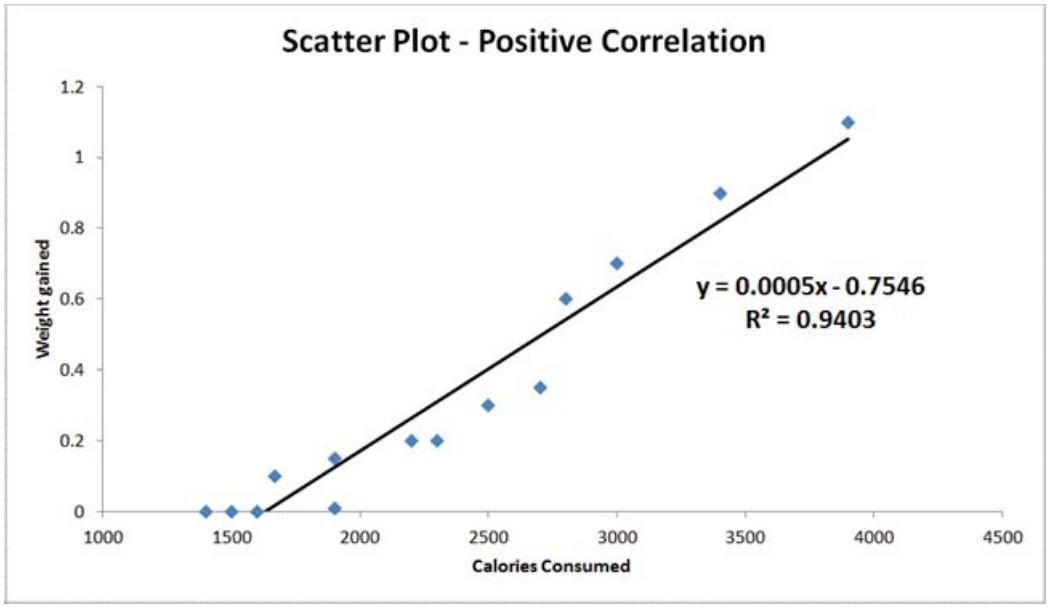
- Scatter plots are the graphs that present the relationship between two variables in a data-set.
- The independent variable is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.
- Also known as scatter graphs or scatter diagrams.

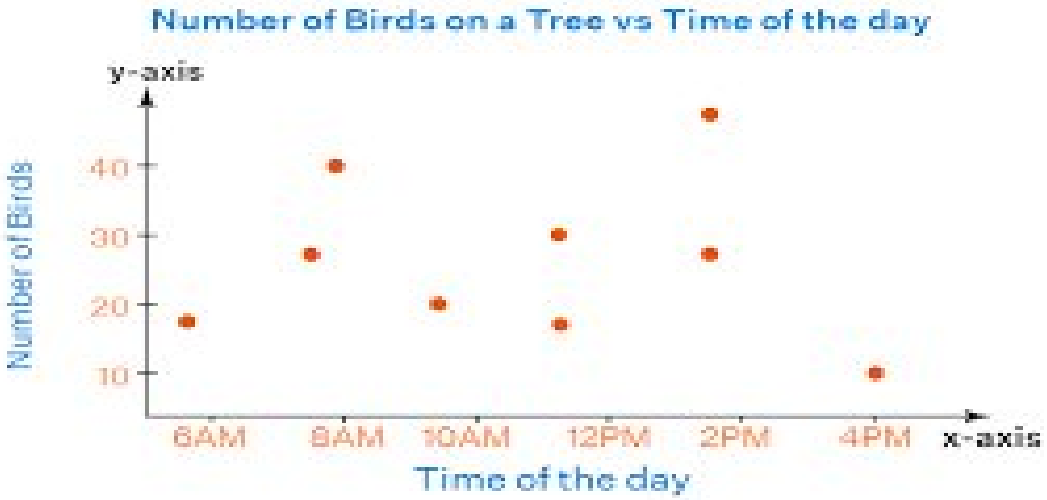
Example:

- Height vs. Weight
- Age vs Height
- Calories consumed vs weight gained
- No. of the birds on the tree vs Time of the day


weight
Height


cuemath
THE MATH EXPERT
Height (in feet)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
2
3
4
5
6
7
8
9
10
11
Age of the child (in years)
Height


Scatter Plot - Positive Correlation
Weight gained
Calories Consumed
y = 0.0005x - 0.7546
R² = 0.9403
1.2
1
0.8
0.6
0.4
0.2
0
1000
1500
2000
2500
3000
3500
4000
4500


Number of Birds on a Tree vs Time of the day
y-axis
Number of Birds
40
30
20
10
6AM
8AM
10AM
12PM
2PM
4PM
x-axis
Time of the day

Identify which graph is best suited for the given scenario

- Sales of different mobile phone brands
- Distribution of students' test scores
- Monthly expenses on rent, groceries, and entertainment
- Temperature change over a week
- Number of pets owned by families
- Heights of students in a class
- lung capacity that how long that person could hold their breath

Exercise 1

The following data shows the number of students choosing different subjects:

Subject	No. of Students
Math	12
Science	18
History	10
English	15

Task

- Draw a ***respective*** graph that representing this data
- Write a programming code for the given scenario.

Exercise 2

The following are the ages of 20 people in a neighborhood:
5, 8, 12, 15, 18, 19, 20, 22, 25, 27, 30, 31, 35, 36, 38, 40, 42, 45, 48, 50

Task

- Create a ***respective*** graph by dividing the ages into bins of 10 (e.g., 0-10, 11-20, etc.).
- Write a programming code for the given scenario.

Measures of Central Tendency (Mean, Median, Mode)

Week 2
Lecture 2
Spring 2025
Rida Maryam

Introduction to Central Tendency

- Represents the center of a dataset.
- Helps summarize large data into a single value.
- Includes: Mean, Median, and Mode.

Example:

Average test scores, median salary, most common shoe size.

Why is it Important?

- Helps in understanding the overall pattern of data.
- Used in decision-making (e.g., business sales, student grades, salaries, etc.).
- Commonly used in research, economics, statistics, and machine learning.

Mean (Arithmetic Average)

- Sum of all values divided by total number of values.
- Affected by extreme values (outliers).

Formula: Mean = (Sum of values) / (Number of values)

- Mean= Σ (x) / N

Example:
A teacher records the scores of 5 students in a math test:
Scores: 85, 90, 78, 92, 88
Mean = 86.6
Interpretation: The average score is **86.6**, meaning most students scored around this value.

Median (Middle Value)

- Middle value of an ordered dataset.
- Less affected by outliers.
- If there is an **odd number** of values, the median is the middle one.
- If there is an **even number** of values, the median is the average of the two middle values.

Example:

Numbers = 5, 12, 18, 22, 30 → Median = 18

Real-Life Example:

- **Salaries in a company:** If a company has extreme salaries (some very high, some very low), the **median** salary gives a better idea of a typical employee's salary than the mean.
- **Household income:** Median income is often used instead of the mean to avoid distortion from extremely high incomes.

Mode (Most Frequent Value)

- Value that appears most often.
- A dataset may have no mode, one mode (Unimodal), two mode(Bimodal) or multiple modes(Multimodal).

- Example: 4, 6, 9, 4, 6, 7, 6 → Mode = 6

Real-Life Example:

- Most common shoe size in a store
- Most frequently sold product in a shop
- Most popular programming language in a survey

Comparing Mean, Median, and Mode

Measure	Best Used For	Limitations
Mean	When data is evenly distributed	Affected by extreme values (outliers)
Median	When data has extreme values or skewed distribution	Doesn't consider all values
Mode	When finding the most common value	May not exist or may not be unique

Conclusion:

- Use **mean** when there are no outliers.
- Use **median** when data is skewed.
- Use **mode** for identifying frequent values.

Class Activity: Solve the Problems

1. Dataset: 5, 10, 10, 20, 25, 30, 10, 40
- Find Mean, Median, Mode

2. Employee Salaries: 2500, 2700, 3000, 3500, 50000
- Which measure best represents typical salary?

Standard Deviation, Variance, and Interquartile Range

Week 2
Lecture 3
Spring 2025
Rida Maryam

Introduction to Dispersion Measures

- **Measures of central tendency** (Mean, Median, Mode) summarize data, but they don't show how spread out the data is.
- **Dispersion** (Spread) tells us how much the data varies from the average.
- **Common dispersion measures**:
  - **Variance**: The average squared deviation from the mean.
  - **Standard Deviation**: The square root of variance.
  - **Interquartile Range (IQR):** The range between the 25th and 75th percentiles.

Variance (σ^2 or s^2)

- **Definition**:
  Variance measures the spread of data points around the mean.
- **Formula for Variance:**
- For a **population variance (σ^2)**:

$$\sigma^2 = \Sigma (x_i - \mu)^2 / N$$

- For a **sample variance (s^2)**

$$s^2 = \Sigma (x_i - \bar{x})^2 / (n-1)$$

Where:
x i = Each data value
μ = Population mean (or $\bar{x}$ for sample mean)
N = Total number of values in the population
n = Total number of values in the sample

Example: If data = 4, 8, 6, 5, 10, then Variance ≈ 5.8.

Example:

Consider data: **4, 8, 6, 5, 10**

Formula: $s^2 = \Sigma (x_i - \bar{x})^2 / (n-1)$

Step 1: Find Mean: $\bar{x}$

$\bar{x}$ =4+8+6+5+10/5=6.6

Step 2: Find Squared Differences from Mean : $(x_i - \bar{x})^2$

$(4-6.6)^2$ =6.76,
$(8-6.6)^2$ =1.96,
$(6-6.6)^2$ =0.36,
$(5-6.6)^2$ =2.56,
$(10-6.6)^2$ =11.56

Step 3: Compute Variance (Sample): $s^2 = \Sigma (x_i - \bar{x})^2 / (n-1)$

s^2 =6.76+1.96+0.36+2.56+11.56 / 5-1
=23.2 / 4 = **5.8**

Standard Deviation (σ or s)

Definition:
Standard deviation is the square root of variance, making it easier to interpret as it's in the same unit as the original data.
Formula:

$$\sigma = \sqrt{\sigma^2} \quad \text{or} \quad s = \sqrt{s^2}$$

Example:
From our variance example:

If Variance = 5.8, then Standard Deviation ≈ 2.41.

Interpretation:

- **Low standard deviation** → Data points are close to the mean.
- **High standard deviation** → Data points are widely spread.

Interquartile Range (IQR)

Definition:

Quartiles are the values that divide a list of numerical data into three quarters (Q1, Q2 and Q3.)
Range is the difference between the maximum and the minimum observation of the distribution. **Range = Xmax – Xmin**

IQR: The difference between the upper and lower quartile is known as the **interquartile range**.
Interquartile range = Upper Quartile – Lower Quartile = Q3 – Q1
IQR measures the spread (range) of the middle 50% of the data.
Steps to Compute IQR:

1. **Arrange data in ascending order.**
2. **Find the median**
3. **Find the first quartile (Q1)** – 25th percentile.
4. **Find the third quartile (Q3)** – 75th percentile.
5. **Compute IQR**

Formula: IQR = Q3 - Q1

Example:

Given dataset: **3, 7, 8, 5, 12, 14, 21, 13, 18**

1. **Sorted**: 3, 5, 7, 8, 12, 13, 14, 18, 21
2. **Q1 (25th percentile)** = 6
3. **Q3 (75th percentile)** = 16
4. **IQR** = 16−6 = 10

Interpretation:

- A **large IQR** indicates a wide spread in the middle 50% of data.
- A **small IQR** suggests less variability.

Real-Life Applications

- **Standard deviation**: Used in finance (market risk), quality control, and grading curves.
- **Variance**: Used in physics and statistics to measure fluctuations.
- **IQR**: Used to detect outliers and summarize income distributions

Class Activity: Solve the Problems

1. Find **variance**, **standard deviation**, and **IQR** for:

Data: 6, 9, 12, 15, 18

2. A company's monthly sales (in $1000s) over 6 months

45, 50, 52, 47, 60, 55

Which measure best shows consistency?