Basic Statistical Concepts

Introduction

Importance of Probability Distributions in Statistics and Data Science

- A probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes of a random variable

Uses of Probability Distributions in Data Science

Descriptive Modeling

- Bernoulli: Models binary outcomes (e.g., success/failure, yes/no).
- Binomial: Repeated Bernoulli trials (e.g., number of heads in 10 coin flips).
- Normal: Continuous data with symmetric distribution (e.g., human height).
- Poisson: Count of events in a fixed interval (e.g., calls per hours).
- Exponential: Time between events (e.g., time until next customer arrives).

What is a Probability Distribution?

Layman Definition:- It’s a way of showing all the possible outcomes of something uncertain (rolling dice) and how likely each one is

Basic concepts

It helps us predict, understand risks, and make better decisions in everyday life, business, and data science

Types: Discrete vs Continuous

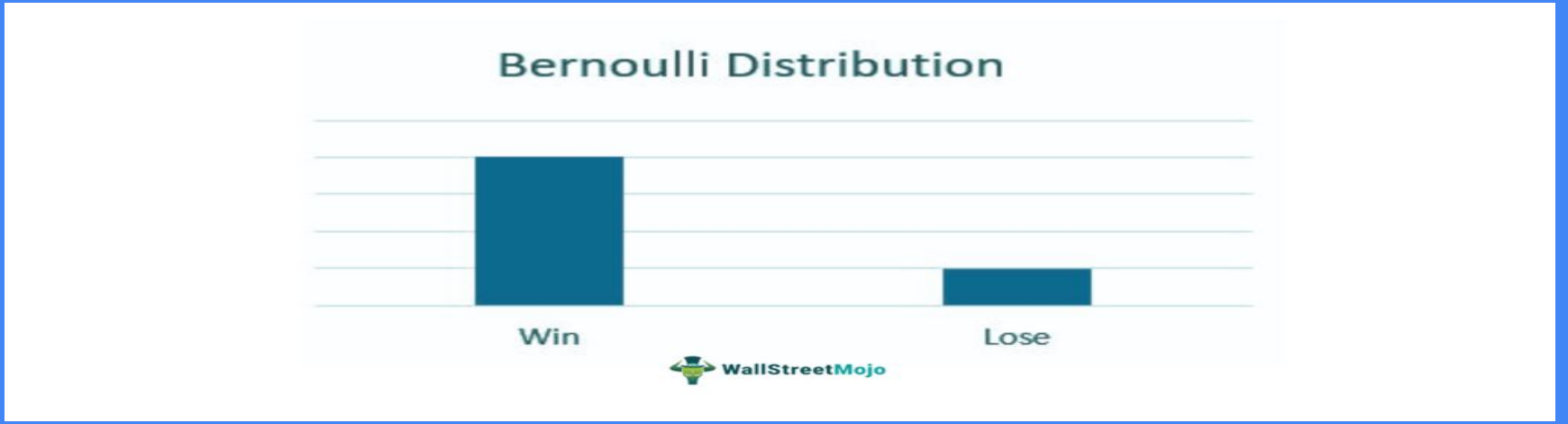
- **Discrete**: Like counting apples — you can have 3 or 4 apples, not 3.5
- **Continuous**: Like measuring water — you can have 2.5 liters, 3.75 liters, etc

Classification of Distributions

- Discrete:- Bernoulli, Binomial, Poisson
- Continuous:- Normal, Exponential

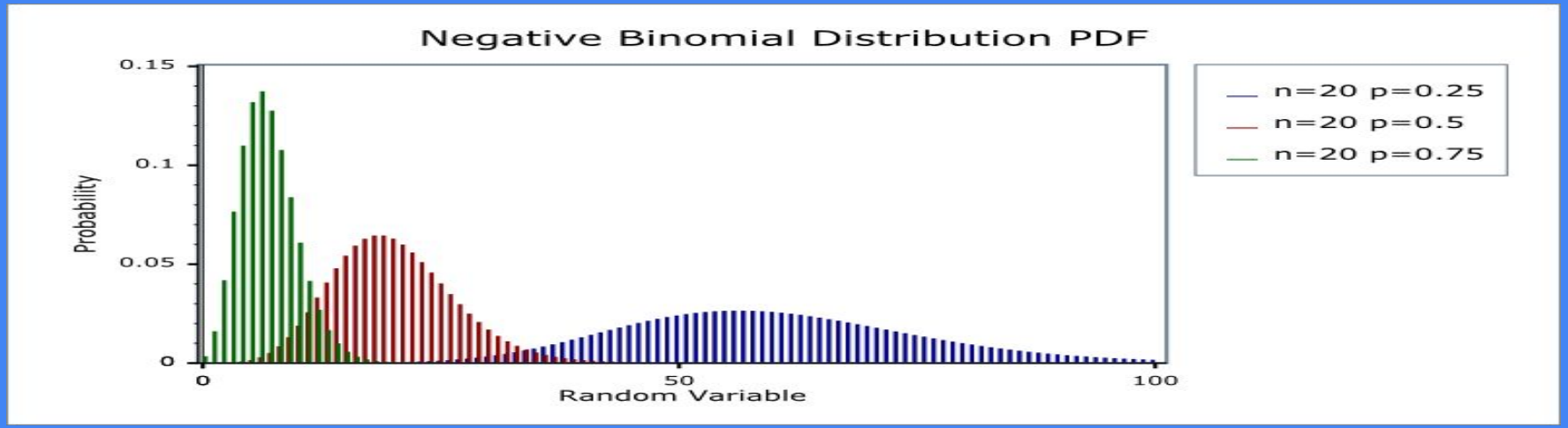
Bernoulli Distribution

- Binary outcome: Success/Failure
- Used in binary experiments



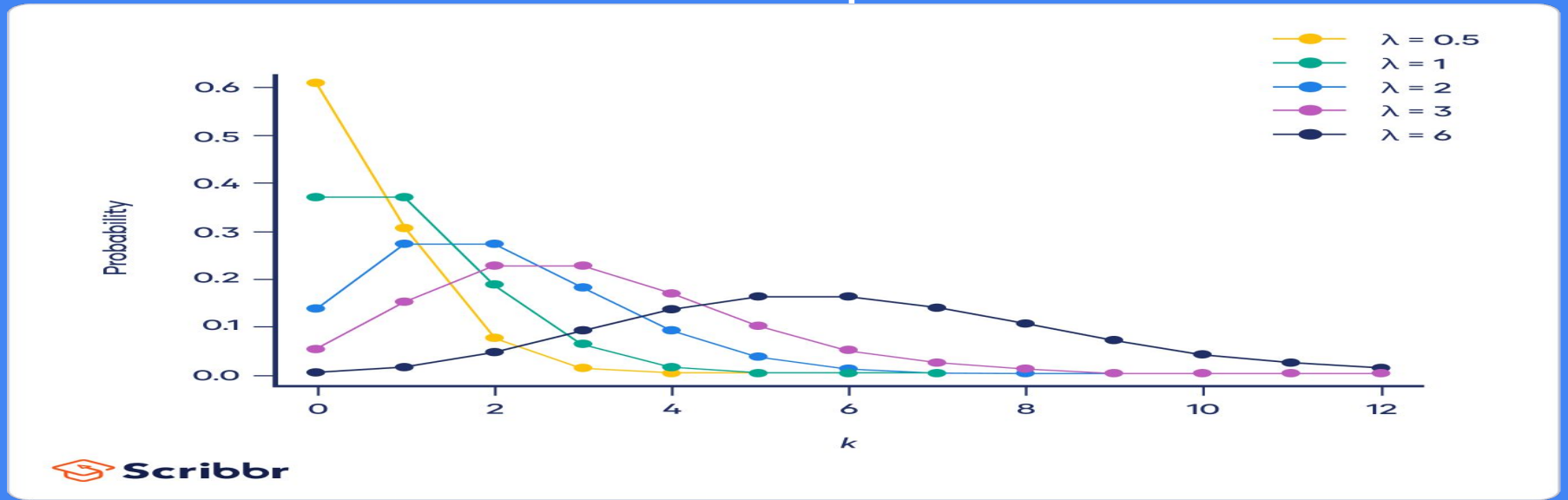
Binomial Distribution

- Multiple Bernoulli trials
- Predicting outcomes over several attempts



Poisson Distribution

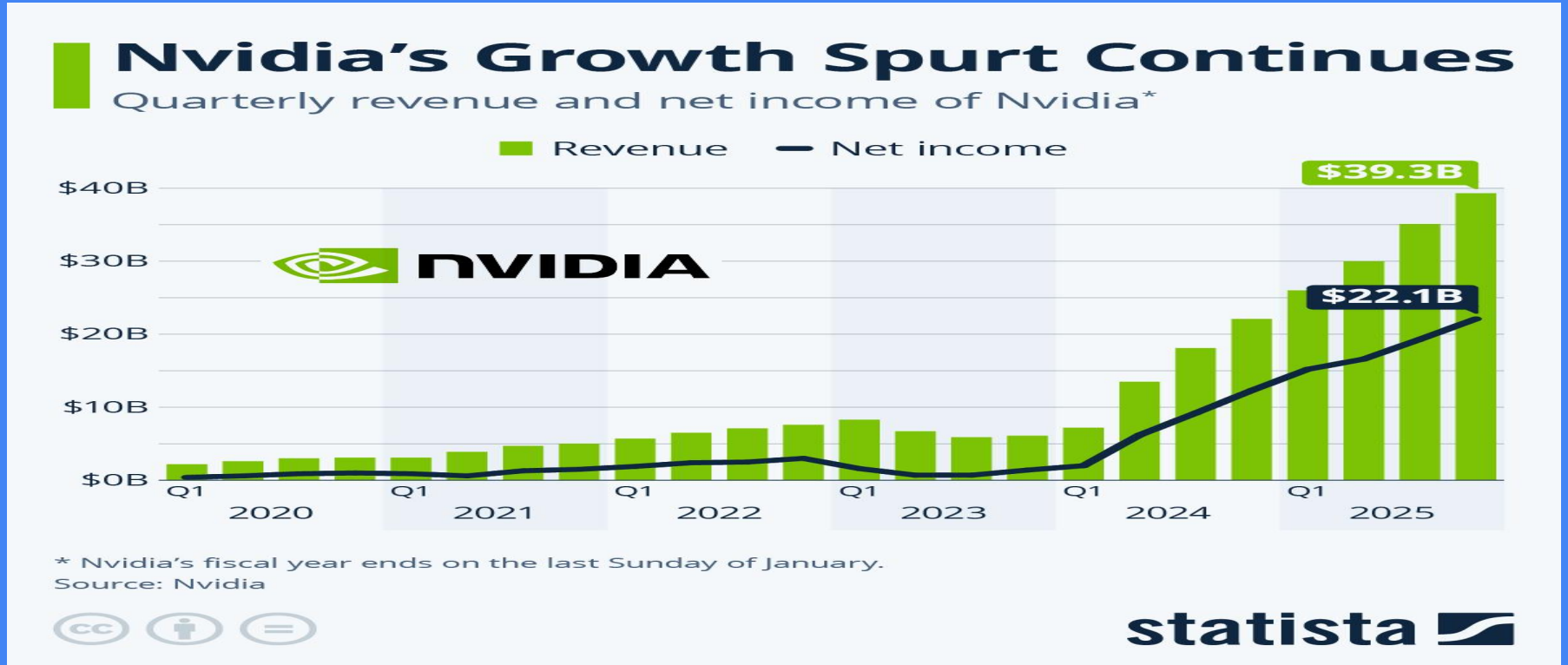
- Call center traffic
- Mean and variance are equal



Exponential Distribution

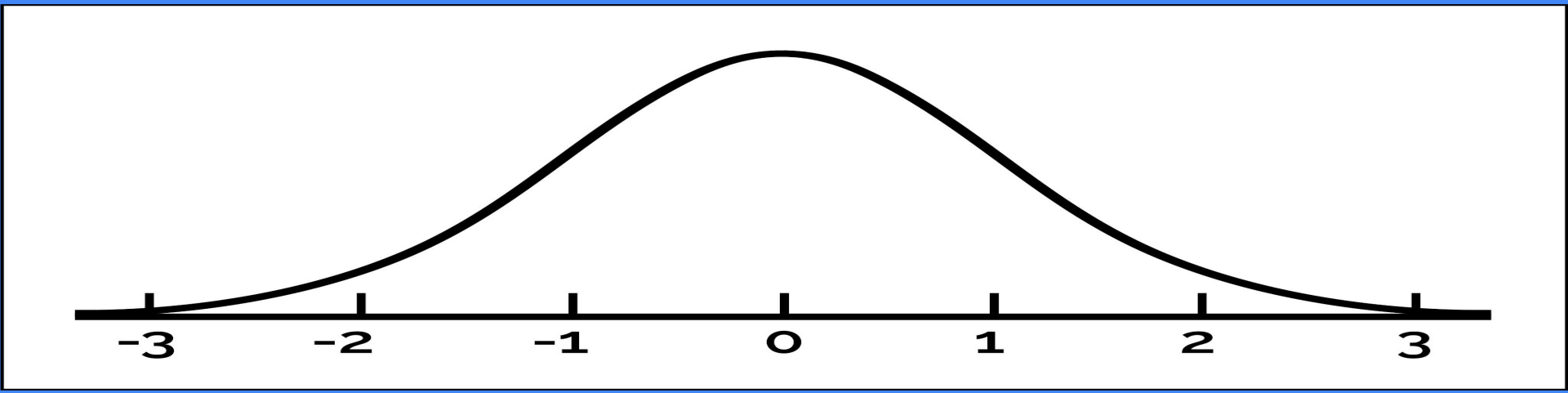
- Time until next customer arrives
- Lifetime of components

Ex....


Nvidia's Growth Spurt Continues
Quarterly revenue and net income of Nvidia*
Revenue
Net income
$40B
$30B
$20B
$10B
$0B
NVIDIA
$39.3B
$22.1B
Q1 2020
Q1 2021
Q1 2022
Q1 2023
Q1 2024
Q1 2025
* Nvidia's fiscal year ends on the last Sunday of January.
Source: Nvidia
statista

Normal Distribution

- Bell-shaped
- Symmetric curve



Python Libraries for Distributions

- libraries like NumPy, SciPy, Matplotlib

Introduction to Correlation

- **Definition:-** Correlation means **how two things move together**

Positive Correlation

- You're observing ice cream sales and temperature.
- On hotter days, ice cream sales go up.
- On colder days, sales go down.
- That’s a positive correlation

Negative Correlation

- Negative correlation means that when one thing increases, the other decreases.
- They move in opposite directions.

Ex

- Sleep vs Stress
- Screen Time vs Eye Health

Zero Correlations

Definition:- Zero correlation means there is **no** relationship between two things

Ex.

- Pizza toppings vs Exam marks
- Hat color vs Monthly salary

How It Works

Key Points:

- Ranges from **-1 to +1**

- **+1** → positive Correlation
- **-1** → negative correlation
- **0** → No Correlations

Importance of Correlation Analysis

Applications:

- Marketing: Ads vs sales
- Finance: Stock price vs interest rate
- Healthcare: Exercise vs blood pressure

Types of Correlation

Positive Correlation

Example:- More study hours ⟶ Higher grades

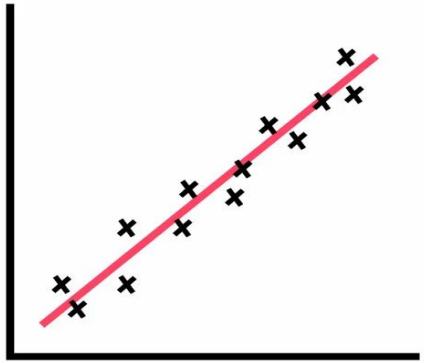
Negative Correlation

Example:- More exercise ⟶ Lower weight

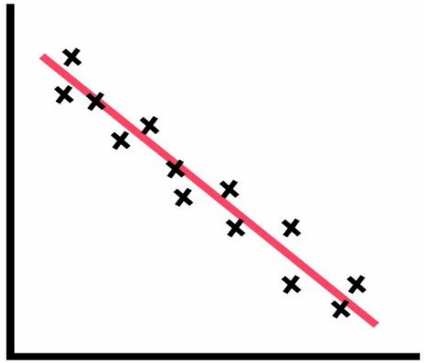
Zero Correlation

Example:- Shoe size and Intelligence

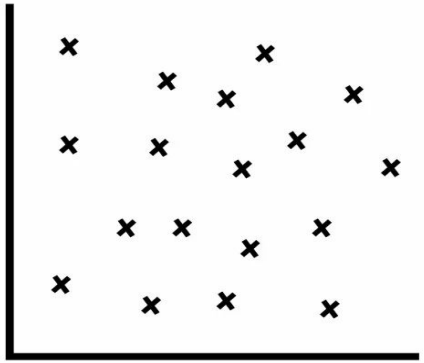
Positive vs Negative Correlation



Positive Correlation



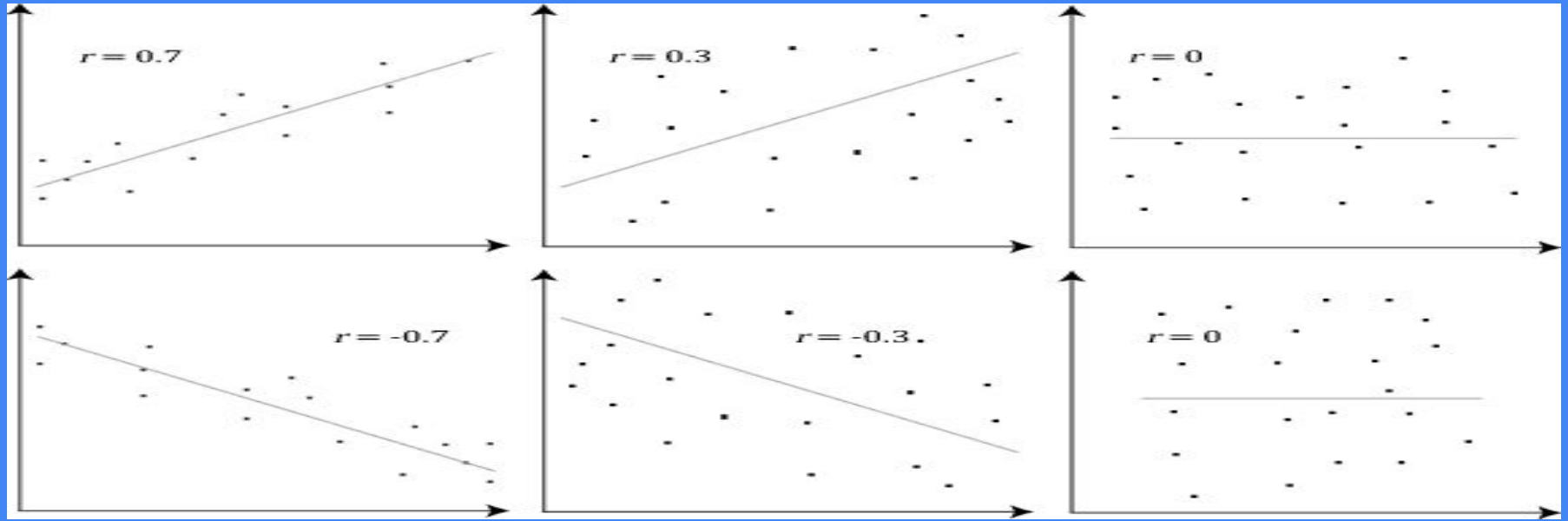
Negative Correlation



No Correlation

Correlation Coefficient (r)

- Also called as Pearson correlation

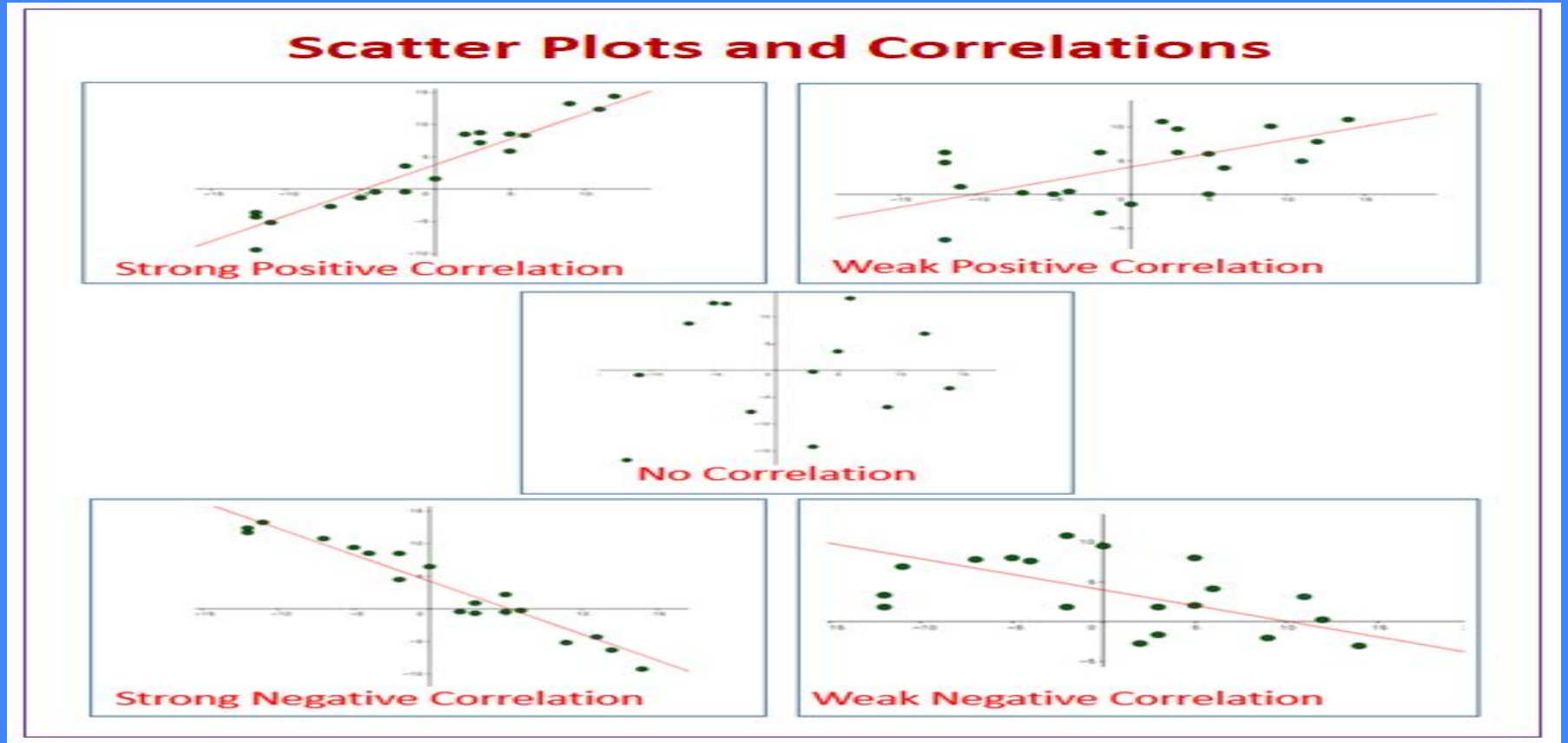


Scatterplots: Definition and Use

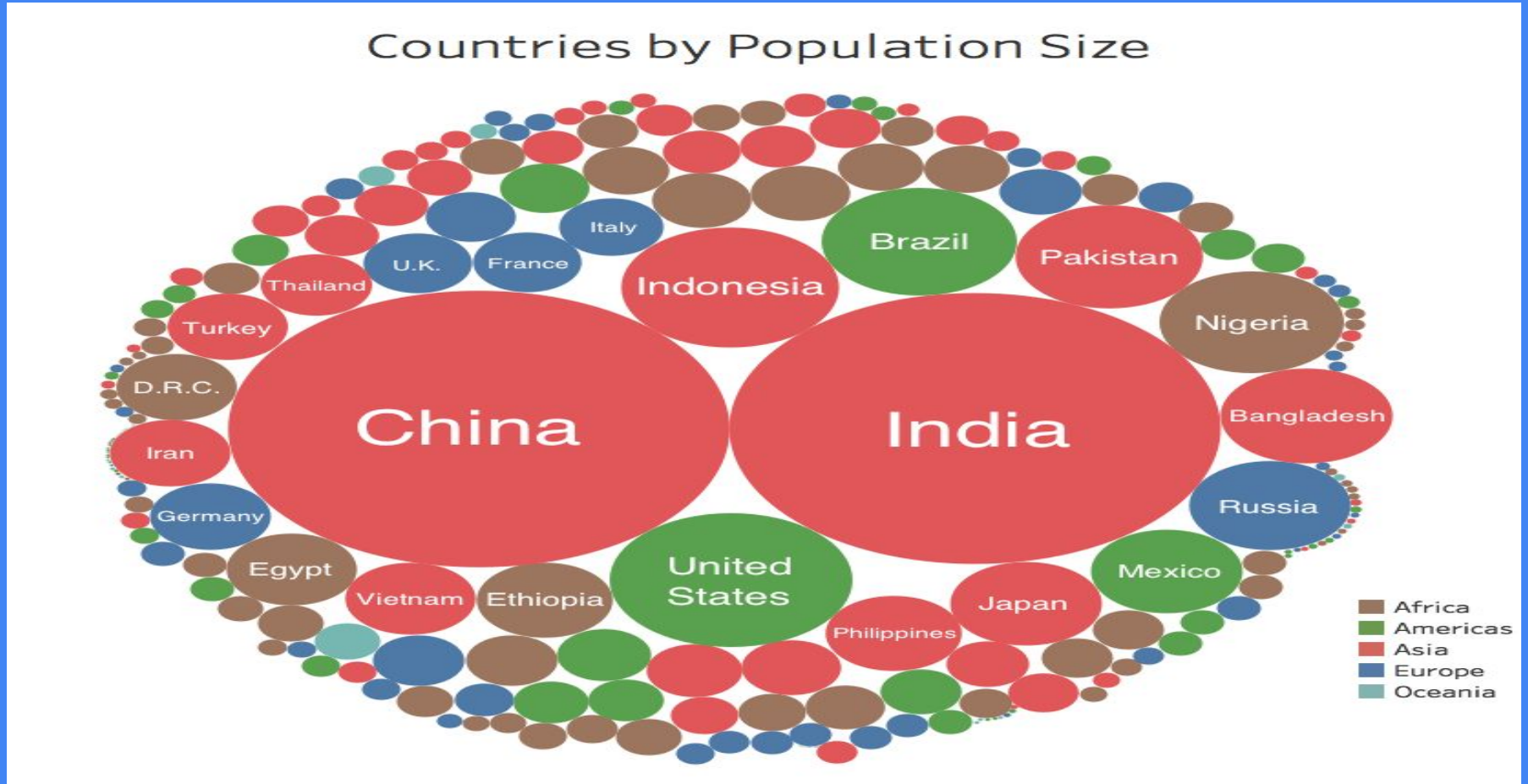
What is a Scatterplot?

- **Def:-** A scatterplot is a graphical representation that shows the relationship between two continuous variables.
- X-axis: Represents the independent variable
- Y-axis: Represents the dependent variable
- Each dot: Represents one data point’s values for both variables

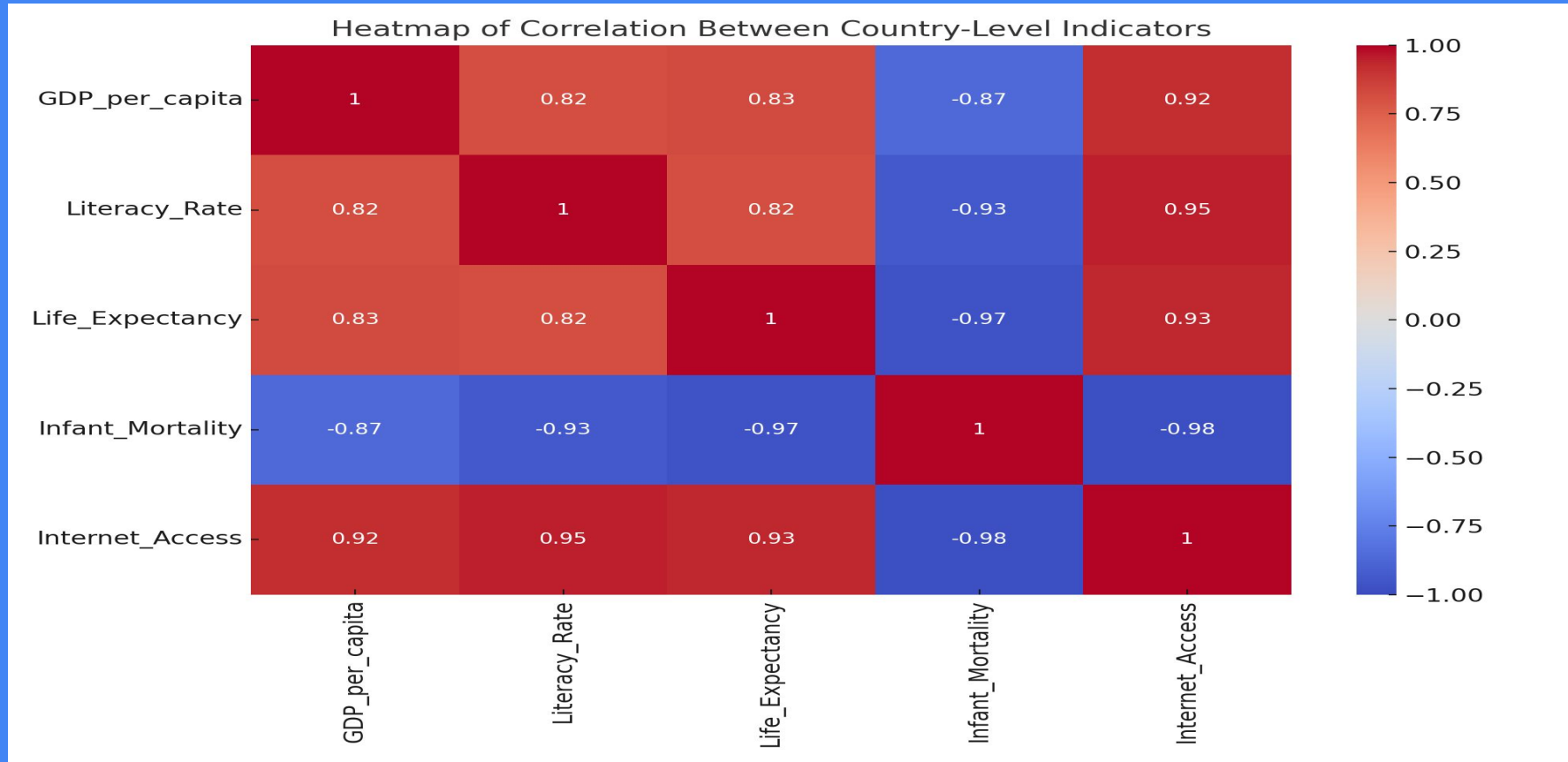
How to Interpret a Scatter Plot



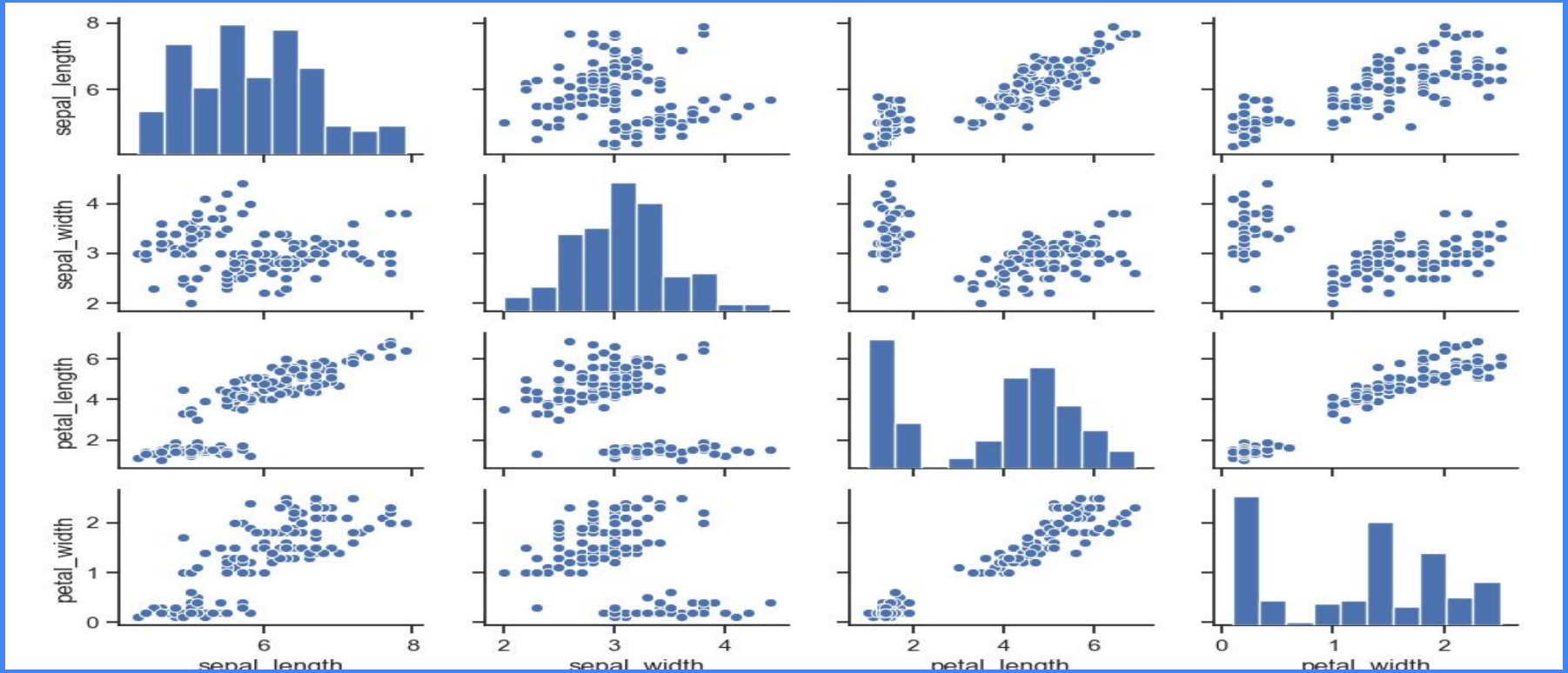
Bubble Charts



Heatmaps for Correlation



Pairplots


sepal_length
sepal_width
petal_length
petal_width
8
6
4
2
3
1
0
sepal_length
sepal_width
petal_length
petal_width