

# SNS COLLEGE OF TECHNOLOGY

Coimbatore-35  
An Autonomous Institution



Department Of Information Technology

## 23ITB302- DATA ANALYTICS

### UNIT-2 DATA CLEANING AND TRANSFORMATION



TOPIC- Data Transformation

A.CATHERINE AP/AIML

# Recap

- Removing Duplicates
- Outlier Detection
- Z-Score, IQR
- Isolation Forest
- Data Cleaning Importance



# Topics Include

- What Is Data Transformation?
- Why Transformation Is Needed
- Types of Data Transformation
- Scaling Techniques
- Encoding Techniques
- Normalization vs Standardization
- Log and Power Transformations

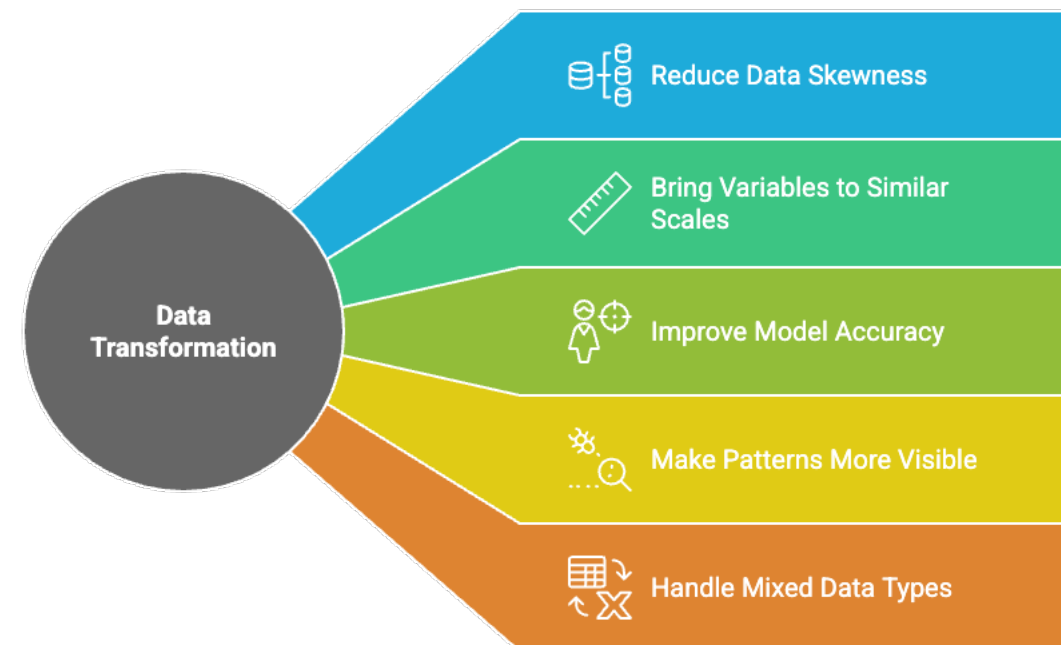
# What Is Data Transformation?

- Converting data from one format/structure/value range to another
- Makes data **machine-readable, consistent, and model-ready**
- Essential for preprocessing in analytics and ML

# Why Transformation Is Needed

- To reduce data skewness
- To bring variables to similar scales
- To improve model accuracy
- To make patterns more visible
- For handling mixed data types

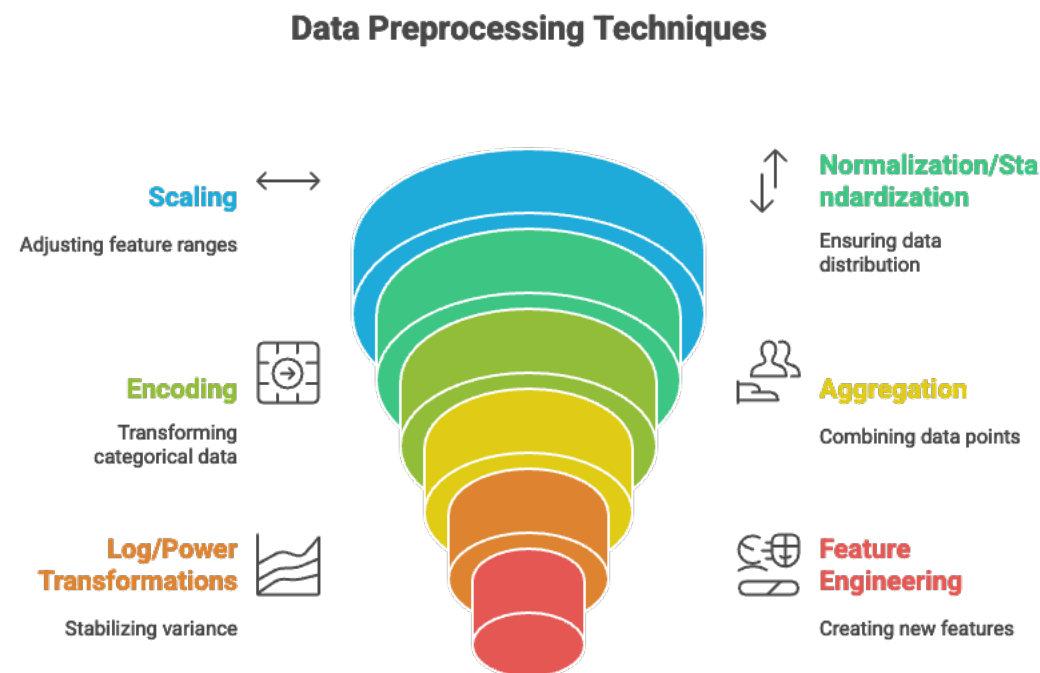
## Unveiling the Benefits of Data Transformation



Made with  Napkin

# Types of Data Transformation

- Scaling
- Normalization & Standardization
- Encoding (for categorical data)
- Aggregation
- Log / Power transformations
- Feature engineering



Made with  Napkin

# Scaling Techniques

- **1. Min–Max Scaling**
- Converts data to a fixed range (usually 0–1)
- Formula:  $(x - \min) / (\max - \min)$
- **2. Standardization (Z-Score Scaling)**
- Mean = 0, Standard deviation = 1
- Formula:  $(x - \mu) / \sigma$

# Encoding Techniques

- **1. Label Encoding**
  - Converts categories into numbers
  - Example: Red=0, Blue=1, Green=2
- **2. One-Hot Encoding**
  - Creates separate columns for each category
  - Example: Male  $\rightarrow$  [1 0], Female  $\rightarrow$  [0 1]
- **3. Ordinal Encoding**
  - Used when categories have order
  - Example: Low=1, Medium=2, High=3

# Normalization vs Standardization

## Normalization

- Values scaled between 0 and 1
- Useful when **data is not normally distributed**

## Standardization

- Centers data around mean
- Useful for **normal distribution** and ML algorithms like SVM, Logistic Regression

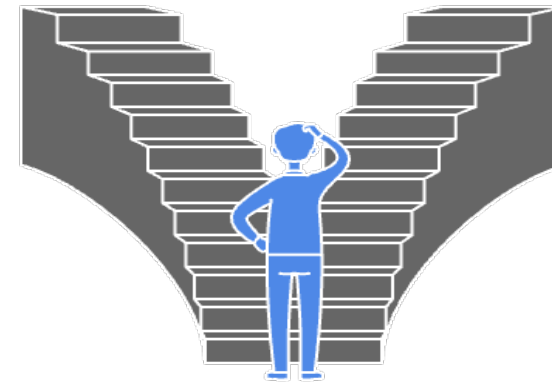
Which scaling method should be used for data preprocessing?

### Normalization

Best for data that is not normally distributed, scaling values between 0 and 1.

### Standardization

Ideal for normally distributed data, centering around the mean.



# Log & Power Transformations

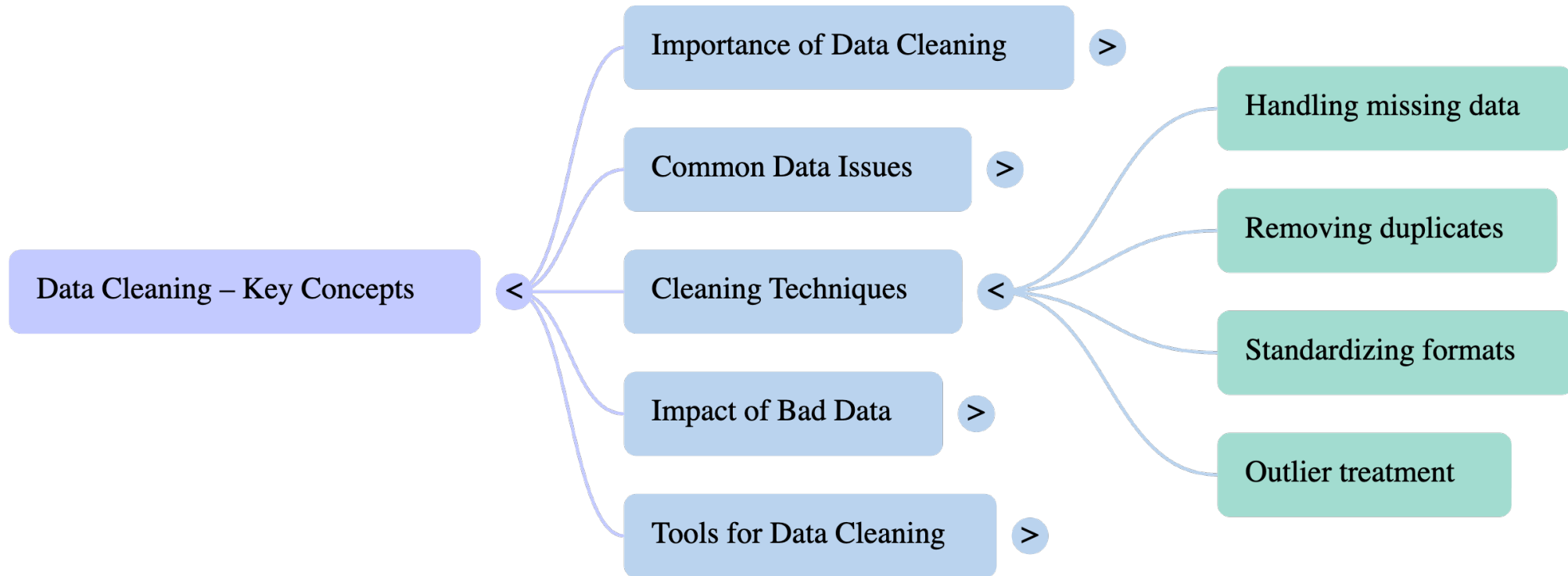
## **Log Transformation**

- Reduces skewness
- Helps stabilize variance
- Good for right-skewed data

## **Power Transformation (Box–Cox / Yeo–Johnson)**

- Makes data more normal
- Useful for modeling

# Mind Map

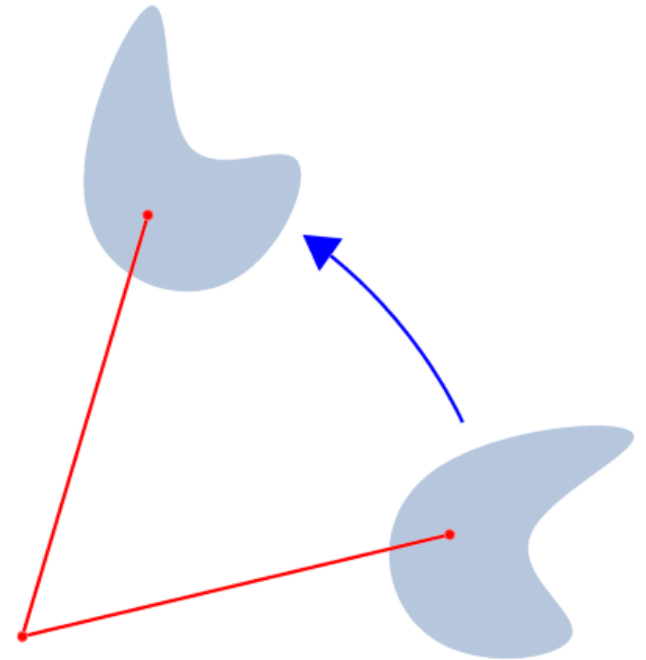


# Conclusion

- Data transformation improves **data quality** and **model performance**
- Different techniques suit different data types
- Essential step before analytics or machine learning
- Ensures consistency, comparability, and interpretability

# Guess the Transformation!

- You have the following situations.  
Choose the best transformation for each:
- Data is extremely right-skewed (income data).
- A numerical column ranges from 1 to 10000, causing model imbalance.
- Categories: Small, Medium, Large (ordered).
- A model requires values to be between 0 and 1.



Thank you!