# EDA (Exploratory Data Analysis) - Notes

## What is EDA (Exploratory Data Analysis)?

What is EDA (Exploratory Data Analysis)?

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics using visualizations and statistical methods.  
  
### Objectives of EDA:  
- Discover patterns and trends in the data  
- Detect anomalies and outliers  
- Test hypotheses  
- Ensure data quality before applying machine learning models.

## Importance of EDA for Machine Learning

Importance of EDA for Machine Learning

Error reading text from shape.

## Understanding Distribution of Data

Understanding Distribution of Data

Data distribution describes how values are spread across a dataset.  
  
### Types of Distributions:

Error reading text from shape.

## Understanding Distribution of Data

Understanding Distribution of Data

## Normal Distribution

Normal Distribution

Shape: Symmetrical, bell-shaped curve.  
Characteristics:  
The mean, median, and mode are all equal and located at the center.  
Represents balanced, natural phenomena (e.g., human heights, exam scores).  
Indication:  
Data is well-balanced without extreme outliers.  
Suitable for applying parametric statistical methods.

Seen in scenarios:  
Linear Regression  
Hypothesis Testing

## Skewed Distribution

Skewed Distribution

Shape: Asymmetrical, with a longer tail on one side.  
Characteristics:  
Occurs when data has extreme outliers.  
Common in income data, sales, or stock prices  
Indication:  
Presence of outliers that could bias the results.  
Data transformations may be needed.

Seen in scenarios:  
Financial Modeling  
Risk Analysis  
Sales Prediction

## Uniform Distribution

Uniform Distribution

Shape: Flat, rectangular shape.  
Characteristics:  
All values have an equal probability of occurrence.  
There is no central tendency (mean and median are less meaningful).  
Indication:  
Data is uniformly distributed, often in simulation or experimental scenarios.

Seen in scenarios:  
Random Sampling  
Simulations  
Monte Carlo Analysis

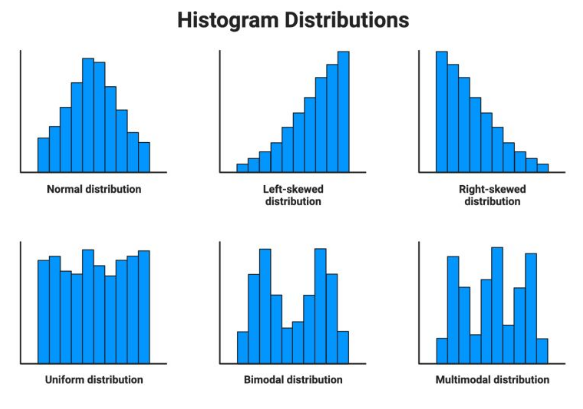
## Bi-modal Distribution

Bi-modal Distribution

Shape: Two distinct peaks   
Characteristics:  
Represents two different groups or phenomena in the data.  
Often indicates the presence of mixed or clustered data.  
Indication:  
The data may contain subpopulations or categories.  
Further segmentation or clustering analysis may be needed.

Seen in scenarios:  
Market Segmentation  
Genetic Studies  
Psychological Research

## Understanding Distribution of Data



## Univariate Analysis

Univariate Analysis

Univariate analysis involves analyzing one variable at a time.  
  
### Methods of Univariate Analysis:  
- Histograms  
- Box Plots  
- KDE (Kernel Density Estimation)  
- Summary Statistics (Mean, Median, Mode)

## Bi-variate Analysis

Bi-variate Analysis

Bi-variate analysis examines the relationship between two variables.  
  
### Methods of Bi-variate Analysis:  
- Scatter Plots  
- Correlation Matrix  
- Heatmaps  
- Line Plots  
  
Used to identify patterns and trends between variables.

## Categorical Analysis

Categorical Analysis

Categorical analysis is used for analyzing variables with distinct categories.  
  
### Techniques for Categorical Analysis:  
- Count Plots  
- Bar Charts  
- Pie Charts  
- Grouped Summary Statistics  
  
Helpful for understanding distribution and frequency of categorical data.

## Correlation and Its Importance

Correlation and Its Importance

Correlation measures the strength and direction of the relationship between two variables.  
  
### Types of Correlation:  
- \*\*Positive Correlation\*\*: Both variables move in the same direction.  
- \*\*Negative Correlation\*\*: One variable increases while the other decreases.  
- \*\*No Correlation\*\*: No relationship between variables.

## Methods to Find Correlation

Methods to Find Correlation

There are different methods to calculate correlation:  
  
### 1. Pearson Correlation Coefficient  
- Measures linear correlation.  
- Values range from -1 to 1.  
  
### 2. Spearman’s Rank Correlation  
- Non-linear relationship analysis.  
- Best for ordinal or ranked data.

## Methods to Find Correlation

Methods to Find Correlation

### 3. Kendall’s Tau Correlation  
- Suitable for smaller datasets.  
- Used to determine ordinal association.  
  
### Visualization Tools:  
- Heatmaps  
- Scatter Plots with Regression Lines.