*WINTER PROJECT*

DATA ANALYTICS

LECTURE 1 :

EDA exploratory data analysis

EDA : EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data.
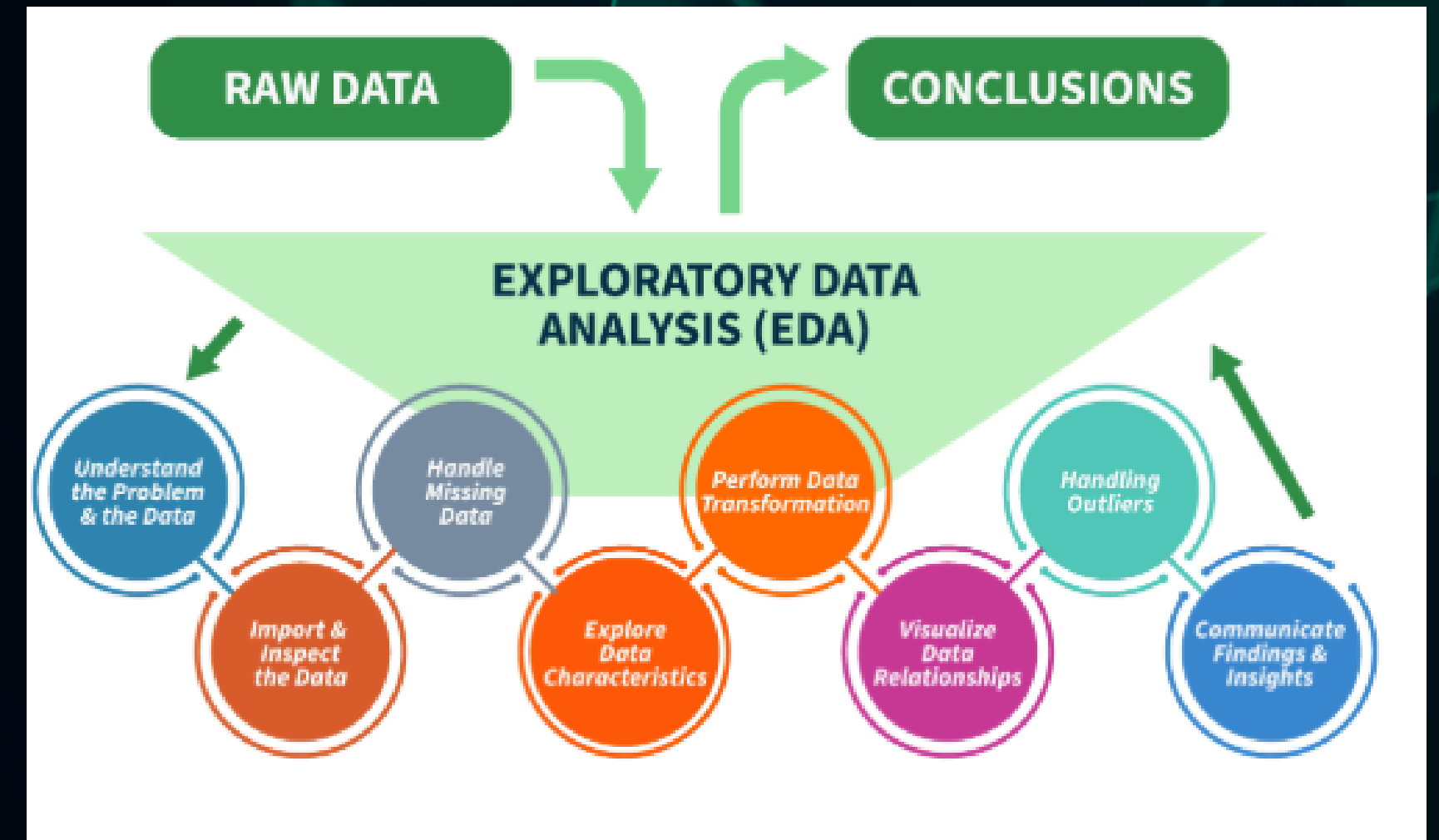
WHY EDA ?

1. **Understanding Data:** Familiarize with features, data types, and distributions for effective analysis.
2. **Revealing Patterns:** Discover relationships and insights for better feature engineering.
3. **Detecting Anomalies:** Identify outliers to prevent skewed results.
4. **Validating Assumptions:** Ensure data meets statistical model requirements.
5. **Feature Optimization:** Select and transform features to improve model performance.
6. **Model Design:** Tailor models based on data characteristics.
7. **Data Cleaning: Spot and resolve missing values or errors for better data quality.**
8. **Clear Communication:** Use visuals and summaries to explain findings effectively.

TYPES OF EDA

- **Univariate:** **Univariate analysis focuses on a single variable to understand its internal structure. It is primarily concerned with describing the data and finding patterns existing in a single feature.**

- **Bivariate: Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variable.**

- **Multivariate : Multivariate analysis examines the relationships between two or more variables in the dataset. It aims to understand how variables interact with one another**
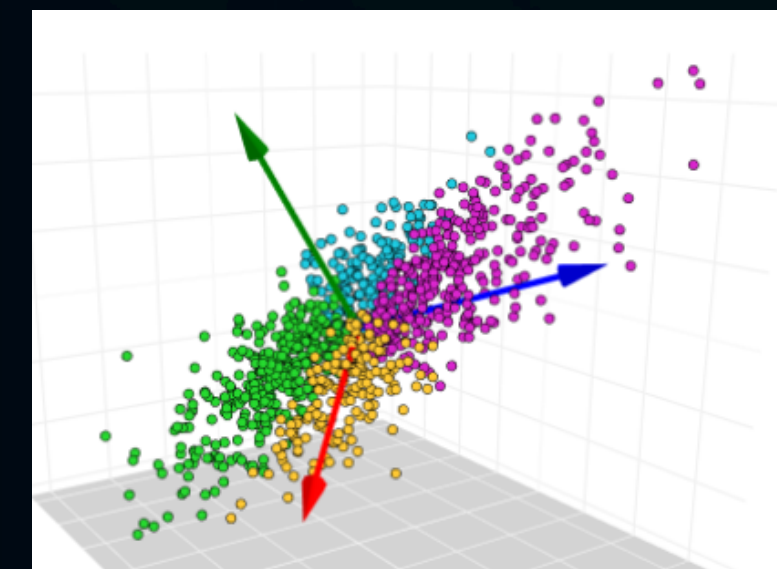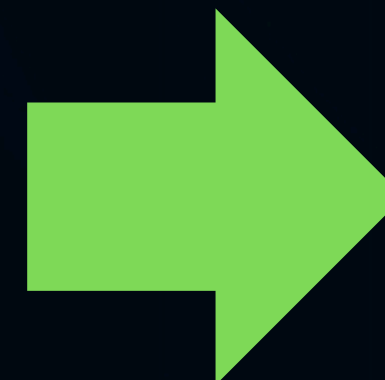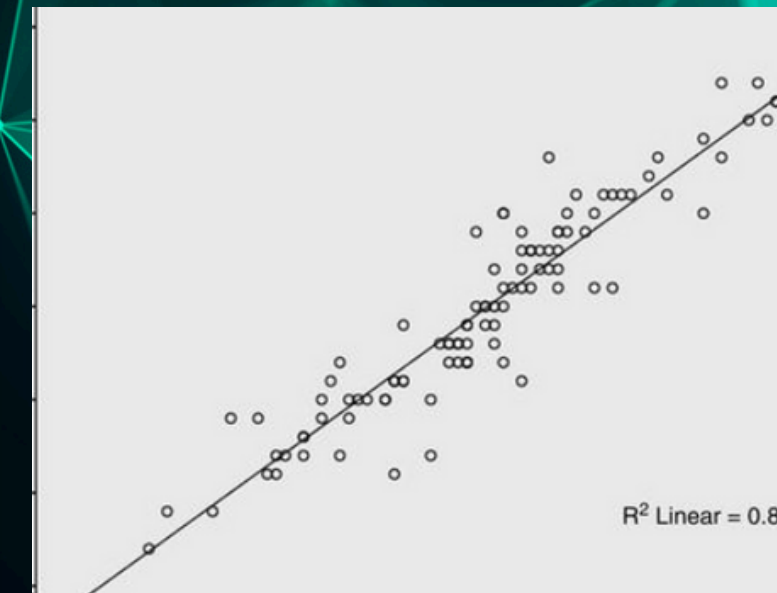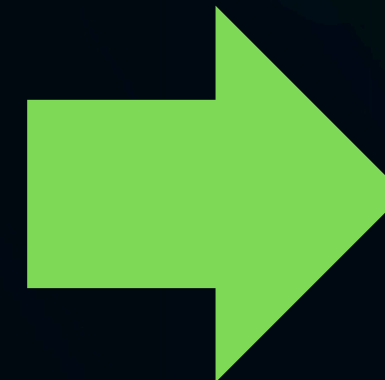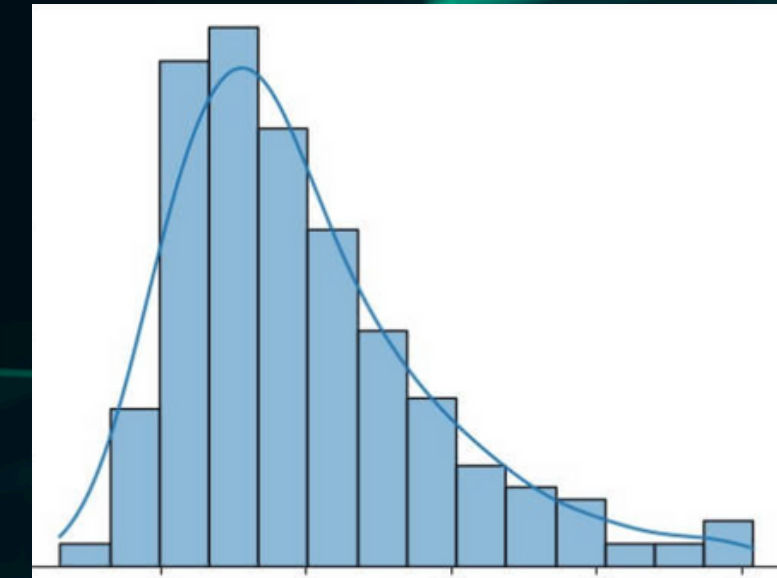
BASIC ANALAYSIS

- loading data
- handle missing values
- handling outliers
- Scalling data

ENCODING

Label Encoding

is a simple and straightforward method that assigns a unique integer to each category. This method is suitable for ordinal data where the order of categories is meaningful.

One hot encoding

One Hot Encoding is a method for converting categorical variables into a binary format. It creates new binary columns (0s and 1s) for each category in the original variable. Each category in the original column is represented as a separate column, where a value of 1 indicates the presence of that category, and 0 indicates its absence.

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding →

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

THANK YOU!