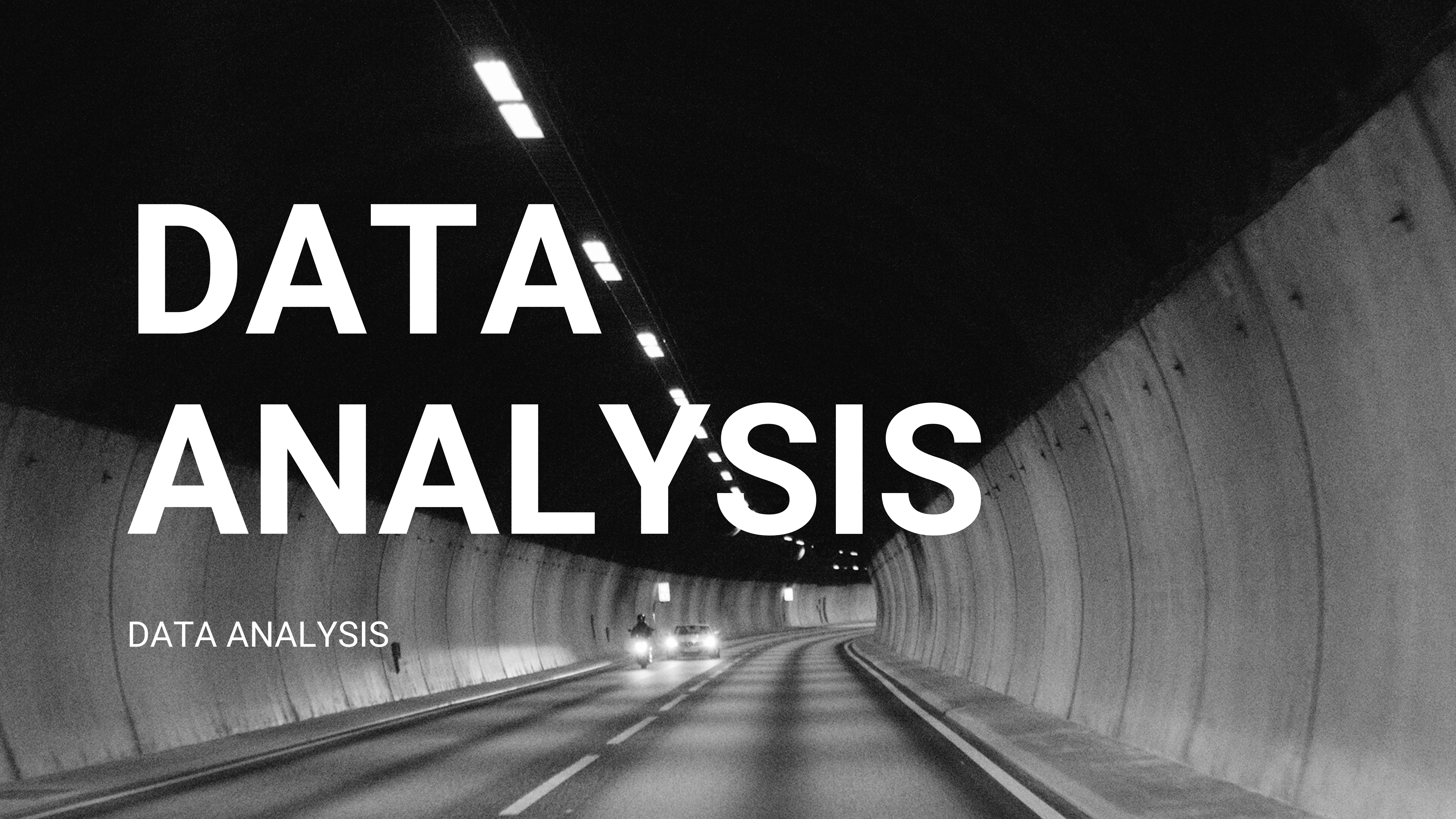

DATA ANALYSIS

DATA ANALYSIS

INTRODUCTION

Data Analysis

Data analysis is the process of inspecting, cleaning, transforming, and interpreting raw data to discover meaningful insights, patterns, trends, and information that can be used to support decision-making, solve problems, or gain a better understanding of a particular phenomenon.

It involves a systematic approach to examine data, which may be in various formats such as numerical, textual, or visual, with the goal of extracting valuable knowledge and actionable conclusions.

Data Analysis Steps

Data Acquisition

The first step is to acquire the raw data from various sources such as CSV files, databases, web APIs, or data scraping. Python libraries like pandas are often used for data ingestion.

Data Cleaning & Preprocessing

After acquiring the data, it needs to be cleaned and preprocessed. This includes handling missing values, removing duplicates, converting data types, and dealing with outliers. Data preprocessing ensures that the data is ready for analysis.

Exploratory Data Analysis

EDA involves exploring and understanding the data. This includes generating summary statistics, creating visualizations (e.g., histograms, scatter plots), and identifying patterns and relationships in the data. Libraries like matplotlib, seaborn, and pandas are used for EDA.

Feature Engineering

In some cases, new features may need to be created from the existing data to better represent the problem. Feature engineering can include scaling, one-hot encoding, or creating interaction features.

Model Building

Depending on the analysis goal, machine learning models may be built to make predictions or classifications. Libraries like sci-kit-learn are commonly used for this purpose.

Model Evaluation

Models need to be evaluated using appropriate metrics to assess their performance. Cross-validation and hyperparameter tuning may be performed to optimize model performance.

Data Visualization

Visualizations, such as plots and charts, are created to communicate the findings effectively. Libraries like Matplotlib, seaborn, and Plotly are used.

Data Cleaning

Data Cleaning & Pre-processing

- Handling Missing Values:
  - Imputation with Mean, Median, or Mode Method: In this approach, missing values are replaced with a statistic computed from the non-missing values in the same feature (column). The most common imputation techniques are:
    - Mean Imputation: Replace missing values with the mean (average) of the non-missing values in the same column. This is suitable for numeric data with a roughly symmetric distribution.

Data Cleaning & Pre-processing

- Handling Missing Values:
    - Imputation with Mean, Median, or Mode Method:
        - Median Imputation: Replace missing values with the median (middle value) of the non-missing values in the same column. This is robust to outliers and is suitable for data with skewed distributions.
        - Mode Imputation: Replace missing values with the mode (most frequent value) of the non-missing values in the same column. This is used for categorical or nominal data.

Data Cleaning & Pre-processing

- Handling Missing Values:
    - Deletion of Rows or Columns Method: In some cases, when the amount of missing data is substantial or when missing data cannot be reasonably imputed, you may choose to remove rows or columns with missing values. This is called data deletion or listwise deletion. Types:
        - Row Deletion (Listwise Deletion): Entire rows containing at least one missing value are removed from the dataset. This approach is suitable when you can afford to lose the incomplete records without significantly affecting the analysis.

Data Cleaning & Pre-processing

- Handling Missing Values:
  - Deletion of Rows or Columns Method:
    - Column Deletion: Entire columns (features) with a large proportion of missing values are removed. This is done when the missing data significantly hinders the usefulness of the feature or when there are too many missing values to impute effectively.

CONCLUSION

That was simple!

QUESTIONS?