Exploratory Data Analysis

- What is EDA?
- Importance of EDA in Data Analysis
- Role of EDA in Data Analysis
- Steps of EDA

What is EDA?

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to:
  - discover patterns,
  - spot anomalies,
  - test hypothesis
  - check assumptions with the help of summary statistics and graphical representations.

Importance of EDA in Data Analysis

- Data Understanding –
  crucial for making informed Decisions about data pre-processing, Modelling & Analysis.
- Data Quality Assessment –
  quality & reliability of data
  Identify missing values, inconsistencies & errors
- Pattern Discovery
  reveal underlying patterns, trends & relationship with data.
- Outlier Detection
  highlight outliers or unusual data points to ensure integrity of data.
- Feature Selection
  identify relevant & informative variables for data analysis/modelling
- Communication
  - Effectively convey complex information to non-technical audiences

Role of EDA in Data Analysis

- **Preparation –**
  Cleaning & Transforming the data
- **Hypothesis Generation –**
  form Hypothesis about the data & its underlying patterns.
- **Visual Exploration**
  Enables to assess data distribution, relationships between variables & presence of outliers.
- **Data Summarization**
  Summary Statistics: measure of central tendency & Variability
- **Data Reporting**
  Generates Insights and findings that are valuable for reporting and decision making

EDA : Data Exploration

- Importance of understanding Data:
  - Contextual understanding – where did data come from? What does each variable represent? What is timeframe of data?
  - Data Quality Assessment – Identify missing information, errors, anomalies etc.
  - Feature Engineering – relevant information or variables.
  - Hypothesis generation – relationships between variables or patterns with data.
  - Data Visualization
- Types of Data:
  - Categorical
  - Numerical

EDA : Data Exploration

- Types of Data:
  - Categorical – represent discrete categories or labels that cannot be measured on numeric scale
    - Nominal Data – no inherent order, example – Colors, types of fruits.
    - Ordinal Data – meaningful order, example – education levels – high school, 10^{th},12^{th}
  - Numerical: numeric & can be measured on continuous or discrete scale.
    - Continuous Data- any number – example – height, temperature.
    - Discrete Data- integers & countable : example – number of customers.

EDA : Data Summary

Techniques used to provide a concise and informative overview of datasets key characteristics to help analyst understand central tendency, variability & distribution of data.

Common techniques

- **Mean** – sum of all values/number of values
- **Median** – middle value in dataset
- **Mode** – most appearing value in dataset
- **Variance** – measures the spread or dispersion of data points around mean.
- **Standard Deviation** – provides average deviation of data points from mean.
- **Range** – difference between maximum and minimum.
- **Percentiles** – helps to understand data distribution and identifying outliers

EDA : Data Visualization

Visually explore, understand and communicate insights of the data.

Common Types of Visualization:

- Histogram – distribution of a single numerical value
- Bar Chart – compare values of different categories or groups
- Line Chart – visualize trends & changes over period of time
- Scatter Plot –relationship & Correlation between two numerical variables.
- Box Plot – distribution of numerical variable – mean, median, quartiles, outliers.
- Heatmap – relationships between two categorical variables
- Pie Chart – represent part of whole.
- Radar Chart – compare multiple quantitative variables for a single observation.

EDA : Data Cleaning & Pre-processing

Identifying & correcting errors, inconsistencies & inaccuracies in data

Common Data Cleaning Tasks:

- Handling Missing Values
- Outlier Detection & Treatment
- Data Standardization & Transformation
- Handling Duplicates
- Encoding Categorical data
- Dealing with Inconsistent data
- Addressing Data Entry Errors
- Handling Skewed Data
- Data Validation & Cross-checking
- Documentation & Versioning