# Data Cleaning and Exploratory Data Analysis (EDA)

## Python -

### Libraries Used

- **pandas** – for data manipulation and cleaning
- **numpy** – for numerical operations
- **matplotlib, seaborn** – for data visualization

## Data Cleaning Process

### General Steps

- Loaded raw .csv files into DataFrames.
- Converted columns to appropriate data types
- Standardized column names for consistency.
- Identified and removed duplicate rows across datasets.
- Handled missing values with suitable imputation strategies.
- Ensured consistency across related datasets (customers, products, sales, stores, returns).

### 1. Customers Data

- **Missing Values:** The age column had 40 null values; imputed with mean age.
- **Feature Engineering:** Created age_group column:
  - < 18 → Youth
  - < 40 → Adult
  - >= 40 → Senior
- **Duplicates:** Found 16 duplicate rows, retained first occurrence, dropped the rest.

### 2. Products Data

- **Missing Values:** The brand column had 60 null values; replaced with "Brand_Unknown".
- **Derived Column:** Created a profit column for each product (based on sales).
- **Duplicates:** Found 24 duplicate rows, dropped them.

### 3. Returns Data

- **Missing Values:** None.
- **Duplicates:** Found 4 duplicate rows, dropped them.

### 4. Sales Data

- **Missing Values:** The store_id column had 992 null values; replaced with "online_store".
- **Duplicates:** Found 60 duplicate rows, dropped them.

### 5. Stores Data

- **Missing Values:** None.
- **Duplicates:** Found 1 duplicate row, dropped it.

## Output

Saved cleaned datasets as:

- customers_cleaned.csv
- products_cleaned.csv
- sales_cleaned.csv
- stores_cleaned.csv
- returns_cleaned.csv

## Exploratory Data Analysis (EDA)

1. **Outlier Detection**
    - Stored DataFrames in a dictionary for iteration.
    - Selected numeric columns across all DataFrames.
    - Generated boxplots for each numeric column.
    - Applied consistent formatting: titles, axis labels, grid lines.
2. **Sales Trend Analysis**
    - Grouped sales data by month to calculate total monthly sales.
    - Converted Period[M] to datetime for compatibility with visualization libraries.
    - Plotted a **time series line chart** of monthly sales.
    - Added grid lines for better readability and visual alignment.

This process ensured **clean, consistent, and reliable data** for further analysis and visualization.