Methodology

This project followed a structured methodology involving data cleaning, transformation, and exploratory data analysis (EDA) to prepare the dataset for meaningful insights and dashboard visualization. The process was executed using **Python (pandas, matplotlib, seaborn)** and involved the following key steps:

---

1. Data Cleaning

The raw datasets contained missing values, duplicates, and inconsistencies that required preprocessing. The cleaning process was applied to each dataset as follows:

1.1 Customers Dataset (customers.csv)

- Converted **signup_date** to datetime format.
- Handled missing values in **age** column by replacing them with the **median age**.
- Converted **age** to integer type.
- Removed duplicate records based on **customer_id**.
- Derived new feature:
    - **age_group** → classified customers as *Teen, Young Adult, Adult, Middle-aged, Senior*.

1.2 Returns Dataset (returns.csv)

- Converted **return_date** to datetime format.
- Dropped rows with missing values.
- Removed duplicates based on **return_id**.

1.3 Sales Dataset (sales_data.csv)

- Converted **order_date** to datetime format.
- Replaced missing **store_id** values with "0".
- Removed duplicate records based on **order_id**.

1.4 Stores Dataset (stores.csv)

- Dropped rows with missing values.
- Removed duplicate records based on **store_id**.

1.5 Products Dataset (products.csv)

- Calculated **sold_units** from sales data.
- Merged with returns data to calculate **return_quantity**.
- Derived new features:
  - **return_rate = return_quantity / sold_units**
  - **profit = unit_price – cost_price**
- Removed duplicate records based on **product_id**.

All cleaned datasets were exported into new CSV files (*_cleaned1.csv) for further use.

---

2. Exploratory Data Analysis (EDA)

EDA was performed on the cleaned datasets to identify patterns, outliers, and key business insights. The following analyses were carried out:

2.1 Outlier Detection

- Used **boxplots** for numeric columns across all datasets.
- Identified extreme values in sales, returns, and product prices.

2.2 Returns Analysis

- Generated descriptive statistics for the returns dataset.
- Verified product-level return trends and abnormal quantities.

2.3 Product Profitability

- Added **profit** column to measure product-level margins.
- Enabled future profitability comparisons across categories and brands.

2.4 Customer Segmentation

- Segmented customers by **age_group** for behavioral insights.
- Enabled targeted analysis of customer sales and loyalty.

2.5 Sales Trend Analysis

- Aggregated monthly sales amounts from **order_date**.
- Created **Monthly Sales Trend line chart** to reveal seasonal sales patterns.

2.6 Top Customer Analysis

- Ranked customers by total sales contribution.

- Visualized the **Top 10 Customers** using a bar chart.

---

3. Summary of Prepared Outputs

- **Cleaned Datasets**: customers_cleaned1.csv, returns_cleaned1.csv, sales_data_cleaned1.csv, stores_cleaned1.csv, products_cleaned1.csv.
- **Derived Features**: age_group, sold_units, return_quantity, return_rate, profit
- **EDA Visuals**: Outlier boxplots, Monthly Sales Trend line chart, Top 10 Customers bar chart.

Concept of df1, df2, df3, df4, df5

In my project, each df corresponds to **one dataset (CSV file)** that I have imported, cleaned, and then enhanced with new features. I have used separate DataFrames (df1, df2, etc.) to keep them organized.