LOW LEVEL DOCUMENT

Exploratory Data Analysis (EDA) Steps

.1. Data Loading and Initial Exploration

- Load the dataset into a pandas DataFrame.
- Explore the first few rows to understand the structure of the data.
- Check the data types, summary statistics, and unique values of each column.
- Identify any obvious data quality issues or inconsistencies.

---

2. Handling Duplicates

- Identify and remove duplicate rows in the dataset.
- **Document**: Number of rows and distinct Order IDs affected by this operation.

---

3. Date Handling

- Normalize the Order Date and Ship Date columns to ensure consistent date formats across all rows.
- Extract the year from the Order ID and compare it with the year in the Order Date. Correct any inconsistencies.
- **Document**: Number of rows and distinct Order IDs affected by these operations.

---

4. Imputation of Missing Values

- **Impute missing values in the Ship Mode column**:
    - Calculate Days to Ship as the difference between Ship Date and Order Date.
    - If Days to Ship is 0, set Ship Mode to "Same Day".
    - If Days to Ship is 7, set Ship Mode to "Standard Class".
- **Impute missing values in the Quantity column**:
    - Choose an appropriate imputation method and **print the rationale** for your selection.
- **Document**: Number of rows and distinct Order IDs affected by these operations.

5. Data Masking and String Handling

- **Drop the Customer Name column**: This is done to protect Personal Identifiable Information (PII).
    - *Note: Protecting PII is crucial for maintaining customer privacy and complying with data protection regulations. Masking or dropping sensitive data like customer names is a critical step in this process.*
- **Create a new column Customer Name Masked**: This column should contain only the initials of the customer name.
- **Convert the Postal Code column**: Change it from numeric to text format, ensuring all codes are 5 characters long. Add a leading '0' where necessary.

6. Data Type Conversion

- Convert the Quantity and Sales Price columns from strings to their appropriate numeric types (int and float, respectively).

7. Handling Inconsistent Categorical Data

- **Clean the State column**: Replace abbreviations with full state names (e.g., "CA" should be changed to "California"). Research state abbreviations online to ensure all entries are corrected consistently.

8. Feature Engineering

Create the following new columns:

- **Original Price**: The price before any discount is applied.
- **Total Sales**: The total revenue generated by multiplying Sales Price by Quantity.
- **Total Profit**: The total profit earned by multiplying Profit by Quantity.
- **Discount Price**: The amount of discount applied, calculated based on Original Price and Discount.
- **Total Discount**: The total discount value for the quantity sold.
- **Shipping Urgency**: Based on Days to Ship:

    - "Immediate" if Days to Ship is 0.
    - "Urgent" if Days to Ship is between 1 and 3.
    - "Standard" if Days to Ship is more than 3.
- **Days Since Last Order**: Calculate days since the last order for each customer.
- **Customer-Level Aggregations**: Create a new dataset storing total sales, quantity, and discount per customer, then merge these back to the original dataset.

---

9. Outlier Detection and Handling

- **Create a function remove_outliers**:
    - This function should take the DataFrame and the column name as arguments.
    - Using the 3 * IQR rule, it should detect and remove outliers and return the cleaned DataFrame.
    - *Why 3*\IQR? The 3 IQR method is applied in situations where the dataset has a high variance, and the standard 1.5 IQR might flag too many points as outliers. This method ensures that only the most extreme values are removed, preserving the integrity of the dataset while still mitigating the influence of true outliers.*
- Use this remove_outliers function to detect and remove outliers from the Sales Price and Profit columns.

---

10. Customer Segmentation and Analysis

- **Calculate Customer Sales Quintile and Customer Profit Quintile**: Based on total sales and total profit per Customer ID.
    - *What is a Quintile? Quintiles are a statistical way of dividing data into five equal parts, each representing 20% of the data. For example, customers in the top quintile (Q5) represent the top 20% of sales or profit.*
- **Create a cross-grid (cross-tabulation)**: Based on these two quintiles to analyze the relationship between customer sales and profitability.