Assignment: Data Cleaning and Analysis using Pandas & NumPy

Objective:

To clean and analyze a *real-world messy dataset* using only **Pandas** and **NumPy**, and generate a detailed analytical report (in text format) explaining meaningful insights.

Dataset Description (Unclean Dataset Provided):

Dataset Description (Unclean Dataset Provided)

The dataset `unclean_sales_data_100.csv` represents **sales transaction records** from a fictional online retail store.
It contains **100+ rows** of data that intentionally include **real-world data quality issues** such as missing values, duplicates, inconsistent formats, and invalid entries.
Students are required to clean, process, and analyze this dataset using **Pandas and NumPy only** (without any data visualization).

Columns Description

Column Name	Description	Common Issues Present
Order_ID	Unique identifier for each sales order.	Duplicate entries, non-sequential order IDs.
Customer_Name	Name of the customer who placed the order.	Inconsistent casing (e.g., “john doe”, “JOHN DOE”), extra spaces, missing values, special characters.
Gender	Gender of the customer.	Mixed cases (Male, MALE, female, Femle), empty or missing values, spelling mistakes.
Age	Age of the customer.	Missing values, negative ages, text values like “thirty”, inconsistent data types.
Product_Category	Category of the product purchased.	Spelling errors (e.g., *clothng*, *Electornics*), inconsistent casing, extra spaces, wrong labels.

Quantity	Number of units purchased.	Stored as strings (e.g., “one”, “two”), missing values, non-numeric data.
Price	Price per unit of the product.	Negative or zero prices, unrealistic values, inconsistent data types.
Total	Total cost of the order (Quantity × Price).	Missing values, incorrect totals, non-numeric entries.
Order_Date	Date when the order was placed.	Multiple date formats (`YYYY-MM-DD`, `DD/MM/YYYY`, `MM-DD-YYYY`, etc.), invalid dates, extra spaces.
Payment_Mode	Mode of payment used by the customer.	Inconsistent casing and spacing (e.g., “Credit Card”, “upi”), missing values, mixed formats.

Part A — Data Cleaning

Tasks

Q1. Load the dataset `unclean_sales_data_100.csv` into a Pandas DataFrame and display:

- The first 10 and last 10 rows.
- The total number of rows and columns.
- Data types of each column using `.info()`.

Q2. Identify and count:

- The number of **missing values** in each column.
- The number of **duplicate records**.
- Drop the exact duplicates from the dataset.

Q3. Clean the **Customer_Name** and **Gender** columns:

- Remove extra spaces and standardize the case (e.g., “John Doe” instead of “ john doe ”).
- Correct inconsistent gender spellings (e.g., “Femle”, “femlae” → “Female”).
- Replace any missing gender values with `"Unknown"`.

Q4. Fix the **Age** column:

- Convert text values like “twenty” or “thirty” into approximate numeric values.
- Replace negative and unrealistic ages with `NaN`.
- Fill missing ages with the mean or median age.

Q5. Clean the **Product_Category** column:

- Remove extra spaces and standardize to proper case (e.g., “Electronics”).
- Correct spelling mistakes such as “clothng”, “eletronics”, etc.
- Replace missing or unrecognized categories with `"Other"`.

Q6. Clean the **Quantity** and **Price** columns:

- Convert non-numeric values in Quantity (like “one”, “two”) into integers.
- Remove or correct negative and zero Price values.
- Convert both columns to proper numeric data types.

Q7. Recalculate the **Total** column:

- Ensure `Total = Quantity × Price`.
- Replace missing or incorrect totals with recalculated values.

Q8. Clean the **Order_Date** column:

- Convert all date formats to a consistent format (e.g., `YYYY-MM-DD`).
- Handle invalid or missing dates appropriately.
- Convert the column to proper datetime type.

Q9. Clean the **Payment_Mode** column:

- Remove extra spaces and standardize names (e.g., “Credit Card” → “Credit Card”, “upi” → “UPI”).
- Replace missing values with `"Unknown"`.

Q10. After cleaning, generate and display the following:

- The shape and info of the cleaned dataset.
- The number of records removed or modified.
- Save the cleaned dataset as **`cleaned_sales_data.csv`**.

Part B — Data Analysis

Tasks

Q1. Display the following basic statistics of the cleaned dataset:

- Total number of records and columns
- Number of unique customers (`Customer_Name`)
- Number of unique product categories (`Product_Category`)

Q2. Compute **customer demographics**:

- Average, minimum, and maximum customer age
- Count of customers by Gender
- Average age of male and female customers separately

Q3. Analyze **sales performance**:

- Total sales revenue (sum of `Total`)
- Average order value
- Highest and lowest order value
- Number of orders with missing or invalid Total before cleaning (you may reuse your cleaned data summary from Part A)

Q4. Determine **top-performing categories**:

- Total sales and number of orders for each `Product_Category`
- Most frequently purchased product category
- Average price per category

Q5. Examine **payment preferences**:

- Frequency (count) of each `Payment_Mode`
- Which payment mode contributes the highest revenue?
- Which payment mode is used least often?

Q6. Perform a **temporal (order date) analysis**:

- Find the earliest and latest order dates
- Calculate the number of orders per month (July, August, etc.)
- Determine average daily sales value

Q7. Identify **data correlations** and quality checks:

- Check the correlation between `Quantity`, `Price`, and `Total`
- Detect and count any outliers (e.g., orders where `Total` > mean + 2 × std)
- Compute the percentage of records that required cleaning in Part A

Q8. Generate a concise **text-based summary report**:
In 3–4 paragraphs, explain your analytical findings in plain English. Include points such as:

- Overall data quality after cleaning
- Key sales insights (total revenue, popular categories, payment modes, etc.)
- Customer age patterns and demographic trends
- Any interesting anomalies or recommendations