## **Dataset Description**

This dataset contains transactional sales data from a multi-national e-commerce company specializing in electronics and office furniture. The company operates across Jordan and Lebanon through multiple sales channels (website, physical stores, etc.) from 2021 to 2025. The dataset includes detailed information about orders, products, customers, sales representatives, and promotional activities.

## **Column Descriptions**

| **Column** | **Description** |
| --- | --- |
| **Order\_ID** | Unique identifier for each order |
| **Date** | Sale transaction date |
| **Customer\_ID** | Unique customer identifier |
| **Country\_Code** | Country where sale occurred (inconsistent format) |
| **Channel** | Sales channel used |
| **Sales\_person** | Name of sales representative |
| **Units\_Sold (quantity)** | Number of units sold |
| **Unit\_Price** | Price per unit |
| **Product\_Code** | Product SKU code |
| **Product** | Product name |
| **Promotion\_Flag** | Whether promotion was active (1=Yes, 0=No) |
| **Commission\_Percent** | Commission rate for sales person |

## **Load and Clean the Data**

### **Basic Cleaning Tasks:**

1. Load the sales datasets **(**sales\_data\_part1.csv, sales\_data\_part2.csv**)** and concatenate them into **one dataframe** using pandas
2. **Rename all column names to snake\_case format** (e.g., 'Order\_ID' → 'order\_id')
3. **Standardize text data, fill missing data with ‘unknown’**
4. **Duplicate Records**

* Identify all duplicate records in the dataset
* Store duplicate records in a separate DataFrame for investigation
* Remove duplicate entries from the main dataset (keep first occurrence)

1. **Handle missing values and anomalies appropriately**
2. **Convert data types** (ensure dates are datetime, numeric columns are numeric)

### **New Columns to Create:**

| **Column Name** | **Description** | **Formula** |
| --- | --- | --- |
| **country** | Standardized country codes | Map country\_code to full names  JO/JOR/Jrdn → **Jordan**  LB/LBN → **Lebanon** |
| **unit\_price\_usd** | All prices in USD | Convert all prices to consistent currency (USD)  Lebanon: no change  Jordan: prices are in JOD, multiply by 1.41or divide by 0.709 |
| **discount\_pct** | Customer discount percentage | The company offers 10% discount to the product if units\_sold > 5 |
| **total\_amount** | Total product value after discount | units\_sold × unit\_price\_usd × (1 - discount\_pct/100) |
| **commission\_amount** | Sales commission earned | total\_amount × commission\_percent |
| **year, month, weekday\_name** | Date components | Extract from date column |
| **sales\_volume\_category** | Volume categorization | 'High' if units\_sold ≥ 5, 'Medium' if 3-4, 'Low' if <3 |
| **customer\_segment** | Customer loyalty tier | Based on total orders per customer: Loyal (≥50), Premium (10-49), Standard (2-9), One-time (1) |

Notes :   
1. Transaction : one record representing an item in an order\_id

2. One order can have multiple products.

## **Business Questions**

1- How many transactions occurred in each sales channel?

2- How many transactions involve each product? Count Product\_Code occurrences.

3- What is the average (mean) unit price across all products?

4- What is the maximum number of units sold in a single transaction?

5- What is the minimum unit price in the dataset?

6- How many sales transactions had a discount?

7- How many unique orders are there?

8- How many customers in each segment?

9- In 2024, which day of the week had the highest sales amount?

10- Identify all sales in 2024 from Jordan where the sales volume is categorized as 'High' and the products sold **were not** 'Gaming Chair', 'Keyboard', 'Monitor', 'Tablet', 'Chair'

* How many sales transactions are there?
* How many sales transactions with ‘Premium’ customer segment?

11- The company is considering closing one of its sales channels in one country due to high operating costs. Using the data, which country and channel combination has the lowest number of transactions, and might be the best candidate for closure?

12- Using the customers.csv file,

* identify the dominant sector among the **loyal** customer segment.
* Which city has the **highest total sales volume (units sold)** in Jordan via the online channel? Show the results sorted in descending order of sales volume.

13- Which products are most frequently sold together in the same order? List the top product pairs and how many times they occurred together?

15-Which day of the week shows the highest sales for each product category? Create a heatmap showing sales volume by product and weekday?

14- Compare weekday sales trends for Jordan vs. Lebanon. Which weekdays are most profitable in each country?

15- Find the top-performing salesperson per year (2022, 2023, 2024) based on total sales amount Show their total sales amount for each year

16- Which salesperson generated the highest commission over the period? Who generated the least?

17- Calculate the Year-over-Year (YoY) percentage change in sales for each country. Visualize the sales trends for all countries over the years. Based on your results, describe how each country is performing. Identify which country shows consistent growth, which is stable, and if any country shows declining trends.

18- Create a monthly summary report for each product showing total sales revenue, number of transactions, and total units sold.

* Show the trend of total sales for a specific product over the months of 2024

19- How many promotion periods were there? For each period, specify the month and year, and list the products included in the promotion?

* Extra : which promotion period increased, decreased, or had no effect on sales of promotions products

## Expected Outcome :

* Every group should present their findings above + need to add around 7 visuals + any recommendations and notes.
* Share your code with us.